6-25-2019

Unified single-cell analysis of testis gene regulation and pathology in five mouse strains

Min Jung

Daniel Wells

Jannette Rusch

Suhaira Ahmad

Jonathan Marchini

See next page for additional authors

Authors

Min Jung, Daniel Wells, Jannette Rusch, Suhaira Ahmad, Jonathan Marchini, Simon R. Myers, and Donald F. Conrad

eLIFE elifesciences.org

RESEARCH ARTICLE

Unified single-cell analysis of testis gene regulation and pathology in five mouse strains

Min Jung[1†], Daniel Wells[2,3†], Jannette Rusch[1], Suhaira Ahmad[1], Jonathan Marchini[2,3], Simon R Myers[2,3*], Donald F Conrad[1,4*]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, United States; [2]The Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; [3]Department of Statistics, University of Oxford, Oxford, United Kingdom; [4]Division of Genetics, Oregon National Primate Research Center, Oregon Health & Science University, Portland, United States

Abstract To fully exploit the potential of single-cell functional genomics in the study of development and disease, robust methods are needed to simplify the analysis of data across samples, time-points and individuals. Here we introduce a model-based factor analysis method, SDA, to analyze a novel 57,600 cell dataset from the testes of wild-type mice and mice with gonadal defects due to disruption of the genes *Mlh3*, *Hormad1*, *Cul4a* or *Cnp*. By jointly analyzing mutant and wild-type cells we decomposed our data into 46 components that identify novel meiotic gene-regulatory programs, mutant-specific pathological processes, and technical effects, and provide a framework for imputation. We identify, de novo, DNA sequence motifs associated with individual components that define temporally varying modes of gene expression control. Analysis of SDA components also led us to identify a rare population of macrophages within the seminiferous tubules of *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ mice, an area typically associated with immune privilege.

DOI: https://doi.org/10.7554/eLife.43966.001

***For correspondence:** myers@stats.ox.ac.uk (SRM); conradon@ohsu.edu (DFC)

[†]These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 32

Received: 28 November 2018
Accepted: 17 June 2019
Published: 25 June 2019

Reviewing editor: Deborah Bourc'his, Institut Curie, France

Introduction

The testis is an amalgamation of somatic cells and germ cells that coordinate a complex set of cellular interactions within the gonad, and between the gonad and the rest of the organism (***Figure 1A***). The key function of the testis is to execute spermatogenesis, a developmental process that operates continually in all male adult mammals in order to generate genetically diverse gametes through recombination and independent assortment of homologous chromosomes. The mechanisms of this process are important for the evolution, fertility and speciation of all sexually reproducing organisms.

A deeper understanding of the transcriptional program of spermatogenesis has potential applications in contraception (***Schultz et al., 2003***), in vitro sperm production for research and the treatment of infertility (***Zhou et al., 2016***), and the diagnosis of infertility, among others. Prior to the advent of single-cell genomics, studies of the highly dynamic transcriptional programs underlying sperm production were limited by the cellular complexity of the testis - comprised of at least seven somatic cell types, and at least 26 morphologically distinct germ cell classes (***Hess and de Franca, 2009***).

The testis has a number of unique features: its transcriptome has by far the largest number of tissue specific genes (over twice as many as the 2nd ranked tissue the cerebral cortex – with which the testis shares an unusual similarity) (***Djureinovic et al., 2014***; ***Guo et al., 2005***; ***Uhlén et al., 2015***); it

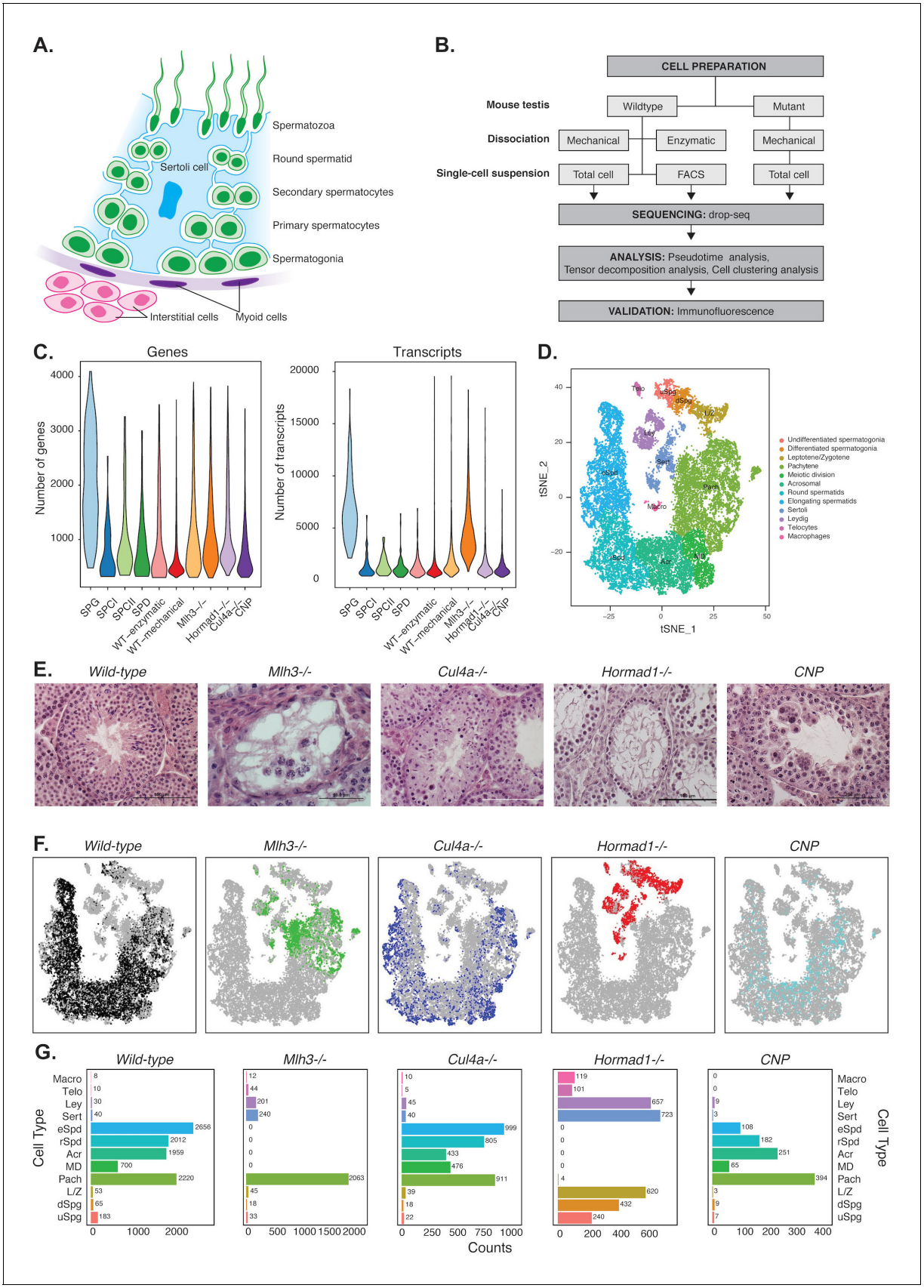

A.
Spermatozoa
Round spermatid
Secondary spermatocytes
Primary spermatocytes
Spermatogonia
Sertoli cell
Interstitial cells
Myoid cells
B.
CELL PREPARATION
Mouse testis
Wildtype
Mutant
Dissociation
Mechanical
Enzymatic
Mechanical
Single-cell suspension
Total cell
FACS
Total cell
SEQUENCING: drop-seq
ANALYSIS: Pseudotime analysis, Tensor decomposition analysis, Cell clustering analysis
VALIDATION: Immunofluorescence
C.
Genes
Transcripts
Number of genes
Number of transcripts
SPG
SPCI
SPCII
SPD
WT-enzymatic
WT-mechanical
Mlh3-/-
Hormad1-/-
Cul4a-/-
CNP
D.
tSNE_1
tSNE_2
Undifferentiated spermatogonia
Differentiated spermatogonia
Leptotene/Zygotene
Pachytene
Meiotic division
Acrosomal
Round spermatids
Elongating spermatids
Sertoli
Leydig
Telocytes
Macrophages
E.
Wild-type
Mlh3-/-
Cul4a-/-
Hormad1-/-
CNP
F.
Wild-type
Mlh3-/-
Cul4a-/-
Hormad1-/-
CNP
G.
Wild-type
Mlh3-/-
Cul4a-/-
Hormad1-/-
CNP
Cell Type
Macro
Telo
Ley
Sert
eSpd
rSpd
Acr
MD
Pach
L/Z
dSpg
uSpg
Counts

Figure 1. Mapping cellular diversity in the adult testis using single-cell expression profiling. (**A**) Anatomy of the testis. Adult testis are comprised of germ cells (spermatogonia, primary spermatocytes, secondary spermatocytes, spermatids and spermatozoa) and somatic cells. Within the seminiferous tubules, there is a population of somatic cells (Sertoli). The tubules are wrapped by muscle-like 'myoid' cells. Outside the tubules are a heterogeneous, poorly defined population of 'interstitial' somatic cells including Leydig cells and telocytes. (**B**) Overview of the experiments. To establish the utility of single-cell profiling for testis phenotyping, we performed a series of experiments (i) comparing the quality of traditional enzymatic dissociation and more rapid mechanical dissociation, (ii) comparing the expression profiles of cells from total testis dissociation to testicular cells of known identity purified by FACS, (iii) comparing expression profiles of wild-type animals to cells isolated from four mutant strains with testis phenotypes (***Figure 1—figure supplement 1***). (**C**) We used Drop-seq to profile 26,200 cells from wild-type animals and 31,400 cells from mutant animals, with an average of 1155 transcripts/cell and 725 genes/cell (wild-type) and 2223 transcripts/cell and 1133 genes/cell (mutants). (**D**) We applied SDA and used t-SNE to visualize cells colored by k-means clustering of 20,322 cells, derived from our full dataset of wild-type and mutant animals, into 32 clusters (Materials and methods, ***Figure 1—figure supplements 1–5***). Label assignment clearly indicates a spatial organization of testis cells in t-SNE space, with somatic cell populations flanking the germ cells in small pockets. The full set of 32 clusters has been simplified into 12 major classes for ease of interpretation; the full clustering is shown in ***Figure 1—figure supplement 2***. (**E**) Histology sections from wild-type and mutant testis, illustrating the phenotypes observed in wild-type and the four mutant strains characterized by Drop-seq. Three of the strains, *Mlh3*$^{-/-}$, *Hormad1*$^{-/-}$ and *Cul4*$^{-/-}$ have known pathology, while strain *CNP* represents an unpublished transgenic line with spontaneous male infertility. (**F**) Mapping of cells from each mouse strain into t-SNE space (colored points) compared to the background of all other strains (gray points). Mutant strains occupy distinct locations within t-SNE space, reflecting the absence of certain cell types in some strains (e.g. *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$), and alteration of expression in remaining cells (e.g. *Hormad1*$^{-/-}$). (**G**) Counting individual cell types provides a quantitative phenotype of cellular heterogeneity in each strain.

DOI: https://doi.org/10.7554/eLife.43966.002

The following figure supplements are available for figure 1:

Figure supplement 1. Comparison of effects of dissociation protocols and mutation status on cell ascertainment and single-cell gene expression.

DOI: https://doi.org/10.7554/eLife.43966.003

Figure supplement 2. Mapping the Cellular Diversity of the Testis.

DOI: https://doi.org/10.7554/eLife.43966.004

Figure supplement 3. Overview of expression patterns for some well known testis cell markers in t-SNE space.

DOI: https://doi.org/10.7554/eLife.43966.005

Figure supplement 4. Tabulation of cluster counts by mouse strain and differential expression analysis within clusters.

DOI: https://doi.org/10.7554/eLife.43966.006

Figure supplement 5. Dissection of Somatic Cell Population Heterogeneity.

DOI: https://doi.org/10.7554/eLife.43966.007

contains the only cells in the male body with sex chromosome inactivation (***Yan and McCarrey, 2009***); meiotic cells undergo programed double strand break formation, homologous chromosome pairing, and recombination; cells undergoing meiosis share transcripts through cytoplasmic bridges (***Braun et al., 1989***); and it features dramatic chromatin remodeling, when the majority of histones are stripped away during spermiogenesis and replaced with small, highly basic proteins known as protamines (***Hammoud et al., 2009***).

Use of genetic tools has enabled the dissection of the homeostatic mechanisms that regulate spermatogenesis, revealing both cell autonomous and non-autonomous mechanisms. However, most perturbations that disrupt spermatogenesis also change the cellular composition of the testis, frustrating the use of high throughput genomic technologies in the study of gonadal defects. By removing heterogeneity as a confounding factor, single cell RNA sequencing (scRNA-seq) promises to revolutionize the study of testis biology. Likewise, it will completely change the way that human testis defects are diagnosed clinically, where testis biopsy is the standard of care for severe cases of male infertility (***Dohle et al., 2012***).

Here, we performed scRNA-seq on 57,600 cells from the mouse testis, using wild-type animals and four mutant lines with defects in sperm production (***Figure 1B***). We set out to develop an analysis approach that would allow us to extract mechanistic insights from joint interrogation of these multiple mouse strains; to gain insights into spermatogenesis and its regulation, using the precise resolution of single-cell analysis; and to establish the utility of scRNA-sequencing for dissecting testis gene regulation in both normal and pathological states.

To do this, we leverage a recently developed factor analysis method, called sparse decomposition of arrays (SDA), which has not previously been applied to single-cell RNA-seq data, and demonstrate how it can be used on scRNA-seq data for cleanup and imputation, identification of co-regulated genes, and to create a dictionary of disease from a joint analysis of mutant and wild-type animals. We show that, unlike standard clustering, we are able to decompose expression patterns into

temporally overlapping yet distinct components, which each possesses specific regulatory mechanisms and functions, providing new insights relative to recent reports of scRNA-seq from mouse testis (**Chen et al., 2018**; **Ernst et al., 2019**; **Green et al., 2018**; **Hermann et al., 2018**; **Lukassen et al., 2018**). Moreover, we retain the ability of other existing scRNA-seq analysis methods to order cells from early to late meiosis, and to identify distinct groups of non-meiotic cells.

Results

Mapping the cellular diversity of the testis with single-cell RNA-seq

To isolate individual cells for data generation, we initially tested two methods for testis dissociation: enzymatic dissociation, a slow 2 hour protocol, vs. a rapid 30 minute protocol based on mechanical disruption (**Lima et al., 2017**). Single cell expression profiles from the two methods showed excellent agreement (r = 0.95), with no important differences in cell quality or ascertainment (**Figure 1**, **Figure 1—figure supplement 1**), so we applied the mechanical dissociation approach for further experiments (**Supplementary file 1**). We performed scRNA-seq to generate 25,423 cell profiles isolated from total testis dissociations of 11 wild-type animals (WT1-WT11). We compared these to reference data for 296 spermatogonia, 199 primary spermatocytes, 398 secondary spermatocytes, and 299 spermatids, purified by FACS (Materials and methods). Transcript yield per wild-type cell (**Figure 1C**, **Supplementary file 2**) were consistent with previous studies using DropSeq on testicular cells (**Green et al., 2018**) or different cell types.

We added to this an additional 31,400 single cell profiles from four different mutant mouse strains exhibiting spermatogenesis defects: three mutants with known molecular mechanisms (knockouts of *Mlh3*, *Hormad1*, and *Cul4a*) as well as one knockin of a transgene (*Cnp*) that led to idiopathic infertility. We performed histological confirmation of testis defects in each animal prior to sequencing (**Figure 1E**). Seminiferous tubules in *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ mice exhibited complete early meiotic arrest and absence of spermatozoa. *Cul4a*$^{-/-}$ sections showed partial impairment of spermatogenesis, with a significant decrease in number of post-meiotic cells and abnormal spermatids. Sections from both *Cul4a*$^{-/-}$ and *Cnp* mice presented giant multinucleated cells, but this type of cell was much more prevalent in *Cnp* seminiferous tubules. *Cnp* mice displayed a clear defect in spermatogenesis; the number of elongating spermatids was grossly reduced to compared to wild-type, and the few elongating spermatids seen in the histology sections featured misshapen nuclear morphology and odd orientation within the disorganized tubules. Sperm tails were occasionally seen in the lumen. Further molecular analysis is required to firmly characterize which stage(s) of spermatogenesis are affected.

Application of SDA, and comparison to classical clustering analysis

One specific challenge of analyzing a developmental system is that cluster-based cell type classification might artificially define hard thresholds in a more continuous process. Furthermore, a single cell's transcriptome is a mixture of multiple transcriptional programs, some of which may be shared across cell types. In order to identify these underlying transcriptional programs themselves rather than discrete cell types we applied SDA (**Hore et al., 2016**). This is a model-based factor analysis method to decompose a (cell by gene expression) matrix into sparse, latent factors, or 'components' that identify co-varying sets of genes which, for example, could arise due to transcription factor binding or batch effects (Materials and methods). Each component is composed of two vectors of scores: one reflecting which genes are active in that component, and the other reflecting the relative activity of the component in each cell, which can vary continuously across cells, negating the need for clustering. This framework provides a unified approach to simultaneously soft cluster cells, identify co-expressed marker genes, and impute noisy gene expression (Materials and methods). We inferred 50 components using SDA. Using these components, we visualized the overall results using t-distributed Stochastic Neighborhood Embedding (t-SNE) (Materials and methods, **Figure 1D**): this t-SNE projection is also used in many subsequent analyses. We estimated the developmental ordering of cells using pseudotime modeling (Materials and methods), based on our t-SNE embedding.

First, to provide a cross-check for our SDA results, we performed k-means (hard) clustering of our single cell libraries into discrete groups. (Materials and methods, **Supplementary file 3**, **Supplementary file 4**). We visualized the resulting 32 distinct clusters in t-SNE space

(Materials and methods, ***Figure 1D***, ***Figure 1—figure supplement 2***). Next, by inspecting the expression levels of known cell type markers and comparing to the FACS-sorted cells, we could resolve our 32 clusters into 11 distinct subtypes of germ cells and four somatic cell populations – Leydig cells, Sertoli cells, immune cells, and telocytes (***Figure 1—figure supplement 2*** and ***Figure 1—figure supplement 3***). By tallying counts of cells within each cluster, we generated a digital readout of the cellular composition of wild-type and mutant animals (***Figure 1G***, ***Figure 1—figure supplement 4***), and are able to associate each SDA component to expression activity in particular cell type(s).

Careful examination and quantification of cell-type composition differences in each mutant strain recapitulated the known pathology of mutants (*Mlh3*$^{-/-}$, *Hormad1*$^{-/-}$ and *Cul4a*$^{-/-}$) at digital resolution. The location of mutant cells in t-SNE space illustrated the absence of certain cell types within spermatogenesis (***Figure 1F***). Consistent with the known biology, we observed that both *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ cells arrest at different stages of meiosis I; mid-pachytene and leptotene/zygotene respectively. Derangement of certain cell types in the developmental trajectory was also observed as leptotene/zygotene *Hormad1*$^{-/-}$ cells formed distinct clusters. Both t-SNE and hard clusters indicated strong mixing of mutant and wild-type cells; of the 32 clusters, only two did not contain both wild-type and mutant cells. Both lacked wild-type cells: cluster 9, a Sertoli cell cluster, and cluster 30, containing leptotene spermatocytes primarily from *Hormad1*$^{-/-}$. As the bulk of our experiments were performed on total testis samples, we do see preferential ascertainment of some cell types from the mutant strains depleted of post-meiotic germ cells: 95% of somatic cells (clusters 1–5,8,9) and 83% of pre-pachytene germ cells (clusters 6, 30–32) are derived from mutants (***Figure 1—figure supplement 4A***). The majority of these clusters have fewer than 10 genes with differential expression detectable between mutant and wild-type (***Figure 1—figure supplement 4B***), and we proceeded with a joint analysis of mutant and wild-type cells, with the caveat that conclusions about the biology of these particular clusters are derived largely from mutant strains.

Following these comparisons, we interpreted our SDA results using published bulk RNA-seq testis data, histology, and GO category analyses.

New molecular markers of cellular subtypes

Single cell RNA sequencing provides new opportunities to assess important open questions in the field of spermatogenesis. Along with the expected patterns of expression for known markers, we identified numerous novel markers for all populations, some of which we selected for validation using immunohistochemistry (***Figure 2***). Noteworthy is the identification of KIF5B as a Sertoli cell protein that provides more extensive coverage of the cell body than the conventional markers TUBB and VIM, and the identification of ABHD5 as a marker for the subcellular structure of developing germ cells known as the residual body. Protein products for predicted markers ACYP1, UNC80, and CCDC62 were not detected, which might be an antibody-related problem or an indication that these RNAs were not translated.

We identified a number of somatic cell populations (hard clusters 1,2,3,4,5, 8 and 9 in ***Figure 1—figure supplement 2A*** 'Merged'). Because our SDA analysis suggested multiple components, varying even within these clusters (see below), we performed additional targeted hard clustering analyses on these cells (Materials and methods), identifying additional complexity: 10 identified somatic cell clusters comprise 4 Sertoli cell sub-clusters, 3 Leydig sub-clusters, two immune cell clusters (macrophages and lymphocytes) and one telocyte cluster (***Figure 1—figure supplement 5***). Telocytes are a recently reported stromal cell type present in a wide range of tissues, and are little studied in testis (***Marini et al., 2018***). In addition to the previously reported markers *Cd34* and *Pdgfra*, we find a number of even more highly specific expression markers for telocytes, including *Dcn*, *Gsn*, *Tcf21* (***Supplementary file 2***).

Each Sertoli sub-cluster is enriched with differing GO terms (biological processes) including cytoskeleton organization (sub-cluster 1), protein folding (sub-cluster 2), RNA splicing (sub-cluster 2 and 3) and spermatogenesis (sub-cluster 4), while Leydig sub-clusters are enriched for steroid and lipid biosynthetic process (sub-cluster 1), ATP synthesis coupled electron transport and drug metabolic process (sub-cluster 2) and cofactor and steroid metabolic process (sub-cluster 3) (***Figure 1—figure supplement 5D***).

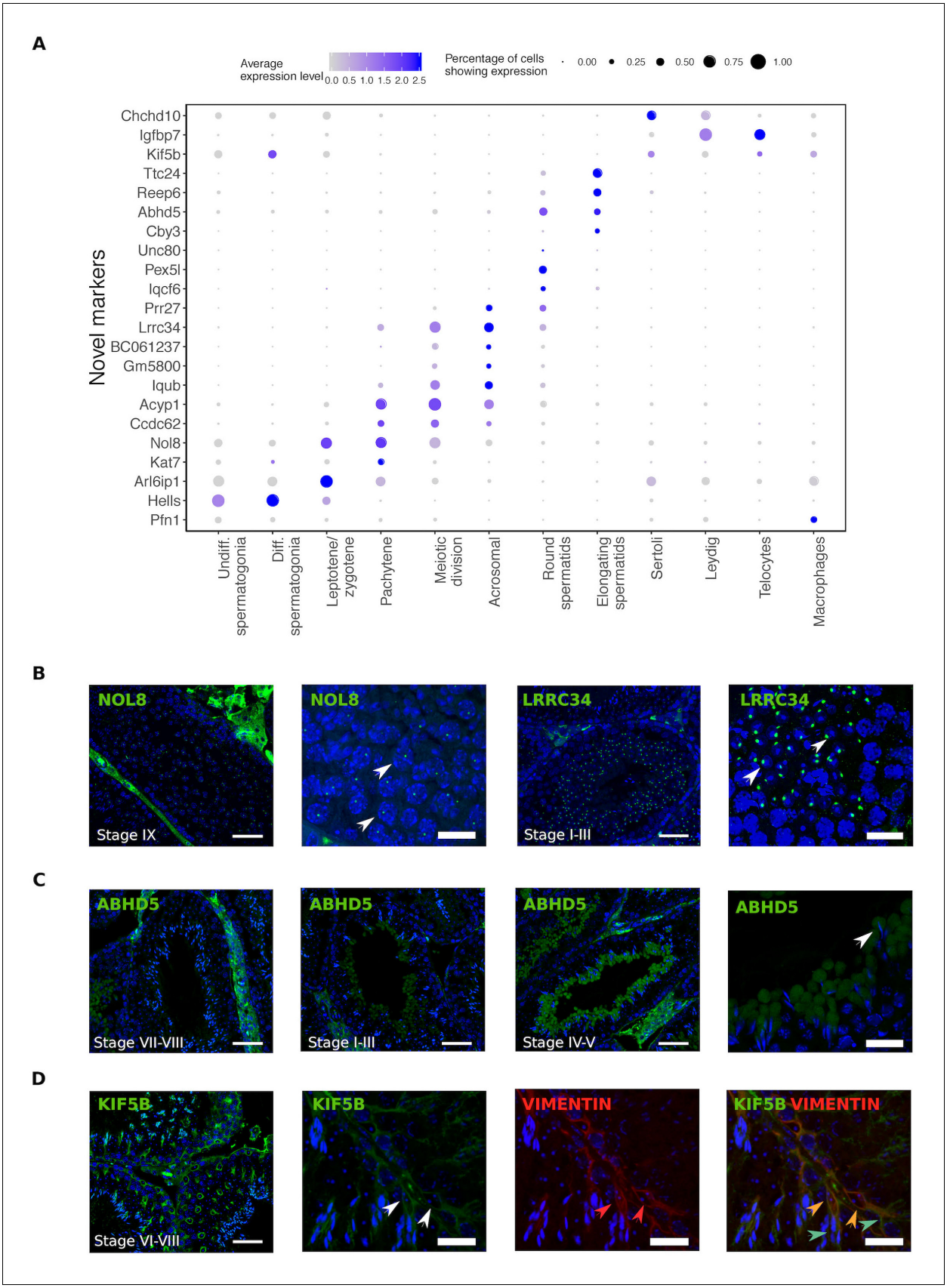

A
Average expression level
0.0 0.5 1.0 1.5 2.0 2.5
Percentage of cells showing expression
0.00 0.25 0.50 0.75 1.00
Novel markers
Chchd10
Igfbp7
Kif5b
Ttc24
Reep6
Abhd5
Cby3
Unc80
Pex5l
Iqcf6
Prr27
Lrrc34
BC061237
Gm5800
Iqub
Acyp1
Ccdc62
Nol8
Kat7
Arl6ip1
Hells
Pfn1
Undiff. spermatogonia
Diff. spermatogonia
Leptotene/ zygotene
Pachytene
Meiotic division
Acrosomal
Round spermatids
Elongating spermatids
Sertoli
Leydig
Telocytes
Macrophages
B
NOL8
Stage IX
NOL8
LRRC34
Stage I-III
LRRC34
C
ABHD5
Stage VII-VIII
ABHD5
Stage I-III
ABHD5
Stage IV-V
ABHD5
D
KIF5B
Stage VI-VIII
KIF5B
VIMENTIN
KIF5B VIMENTIN

Figure 2. Identification of novel cellular markers from single-cell data. (**A**) Across major cell-type clusters, we identified 22 gene expression markers specific to one cell type or aspect of spermatogenesis and not previously reported. Here we show the expression levels of these genes. Expected protein expression patterns for *Nol8*, *Lrrc34*, *Abhd5*, and *Kif5b* were confirmed, but the antibodies for *Acyp1*, *Ccdc62*, and *Unc80* did not show positive staining in any testicular cell types,which could be an antibody-related problem or an indication that these RNAs were not translated. (**B–D**) Thin scale bar, 50 μm; thick scale bar, 20 μm. (**B**) *Nol8*, a nucleolar protein, marks primary spermatocytes while *Lrrc34* marks nucleoli in round spermatids (white arrowheads) (**C**) Within the tubules, *Abhd5* marks specific cytoplasmic regions of elongating spermatids destined to form the residual body (white arrow head) and staining intensity peaks during seminiferous tubule stages IV-V. (**D**) *Kif5b* marks Sertoli cells within seminiferous tubules (white arrow head). We co-stained *Kif5b* with a well-known Sertoli cell marker, Vimentin (red arrow head), and indeed both proteins colocalize to Sertoli cells (orange arrow head). Co-staining also reveals that *Kif5b* staining extends further out in the cell body (blue-green arrow head) than Vimentin.
DOI: https://doi.org/10.7554/eLife.43966.008

SDA-based gene expression modules

Based on the above analyses, it is clear that our 50 SDA components represent all the major different cell types and developmental stages of spermatogenesis, with other specific components capturing batch effects and general processes such as respiration. Encouragingly, most components contained relatively few highly expressed genes (***Figure 3B***) and, when compared to alternative commonly used methods for matrix factorisation (non-negative matrix factorization, NNMF; principal component analysis, PCA; independent component analysis, ICA), SDA produced the most sparse model (***Figure 3C***). Although the model used for inference is symmetric for positive/negative gene weightings, many identified components showed strong biases towards positive or negative weights, consistent with expectations for identifying a group of co-activated (or co-repressed) genes (e.g. ***Figure 3E***). Likewise, the cell loadings of each component frequently highlight specific cellular subsets that localize in t-SNE space and pseudotime (***Figure 3D&F***, ***Figure 3—figure supplement 1***), and often interpretable as particular identifiable cell types in our initial hard clustering. Thus, we label SDA components as 'expression modules'. We found that most components generated from an SDA analysis of only wild-type data were also observed in the joint analysis of wild-type and mutant data, which we proceed to use for the remaining analyses (***Figure 3—figure supplements 2–4***).

To provide further intuition towards how SDA components summarize transcriptional programs, we selected 14 components that, collectively, load highly on germ cells throughout spermatogenesis. When we visualize the total expression output for each cell, ordered by pseudotime, as a sum of all 14 components, it is clear that expression can be modeled as an overlapping series of components in time, coming on and going off gradually over different timescales (***Figure 4A–B***). Each component is enriched for specific genes, and, importantly, genes with different identified biological functions (***Figure 4C&D***). SDA components provide complementary information to hard clustering: a single hard cluster may have significant cell scores from as many as three components, indicating multiple different expression programs jointly active in each cell. Conversely a single SDA component may show significant cell scores across more than three hard clusters (***Supplementary file 3***), emphasizing that expression changes gradually as cell types and fates evolve (***Figure 4—figure supplements 1–2***).

In addition to identifying soft clusters and their markers, by multiplying the cell scores and gene loadings, SDA can impute very sparse, noisy, expression data. In principle, harnessing the correlation structure of gene coexpression across cells can improve predictions, overcoming the sparsity of the initial data. Indeed, our dataset has 93.8% zero values and a median of 1,312 UMI transcripts per cell. Nonetheless, SDA imputation is able to estimate expression of individual genes even when in many cells zero reads are observed (***Figure 5A***). It is not possible to determine the true expression vector for an individual cell, so we use cross-validation to test whether imputation improves expression estimates. Specifically, we assign each read to either a training or test set. We predict gene expression based on the training set, using the SDA approach, or another approach (e.g. the dedicated single cell imputation method MAGIC; ***van Dijk et al., 2018***), and then evaluate our ability to rank gene expression using the test set (Materials and methods). SDA imputation outperforms approaches using the raw data, for essentially all cells in the test data (***Figure 5B&C***, ***Figure 5—figure supplement 1C***). While providing the most sparse representation (***Figure 3C***), SDA still imputes equally well, compared to other matrix factorizations and to MAGIC (***van Dijk et al., 2018***) (***Figure 5C***, ***Figure 5—figure supplement 1A***). Further, when compared to NNMF, SDA provides

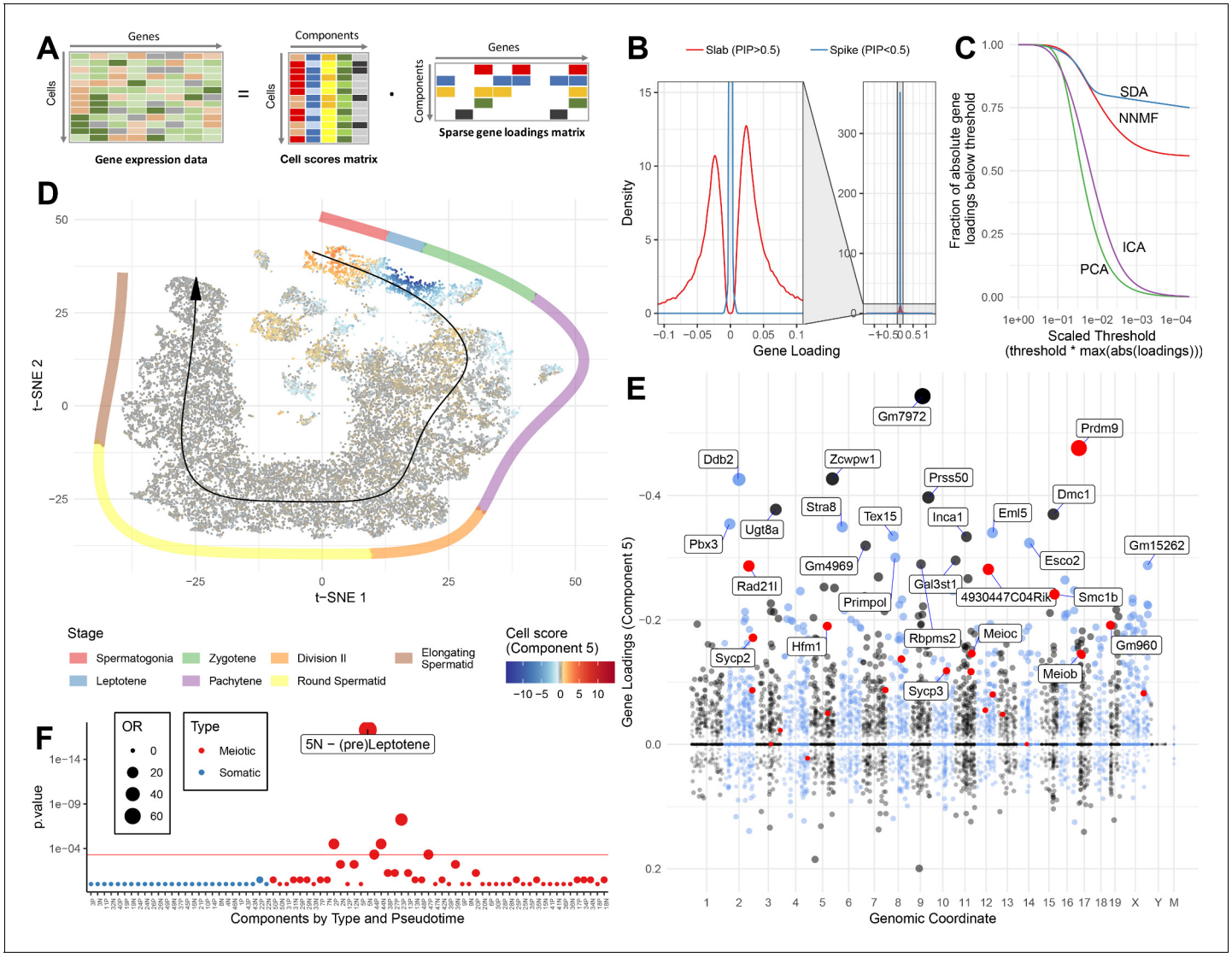


Figure 3. SDA identifies gene modules and maps them to cells. (**A**) We applied sparse decomposition analysis (SDA) to identify latent factors ('components') representing gene modules. These components are defined by two vectors – one that indicates the loading of each cell on the component, and one that indicates the loading of each gene on the component. (**B**), SDA uses a spike and slab prior on the gene loadings to induce sparsity (a point mass at 0 and a centered normal distribution respectively). PIP = Posterior Inclusion Probability that a gene loading is not equal to zero (i.e. not in the spike). The figure shows the density of gene loadings over all components with loadings separated into genes with PIPs > 0.5 (20%) versus <0.5, indicating the sparsity of resulting gene loadings. (**C**) SDA produces sparser representation of gene loadings compared to other matrix factorizations: NNMF, ICA and PCA. For each method, the fraction of all absolute gene loadings exceeding a 'no loading' sparsity threshold is shown, normalized by the maximum absolute loading across all components for that method. (**D**) We fitted 50 SDA components using 20,322 wild-type and KO cells (see also ***Figure 3—figure supplements 1–5***). We illustrate component 5. The loadings of component 5 in t-SNE space highlight a cluster of cells at the leptotene early meiotic developmental stage. Black arrow: the principle curve fit to the germ cell data, corresponding to the developmental ordering of each cell progressing through spermatogenesis. The colored segmented line shows broad staging of spermatogenesis. (**E**) Genomic location versus loadings for component 5. Most genes have near-zero loadings, but a fraction have non-zero loadings, including the well-known histone methyltransferase *Prdm9*. Red genes: GWAS hits for human recombination rate. (**F**) Component 5 is highly and specifically enriched for GWAS hits of human recombination rate. OR: Odds Ratio. P value by FET (main text). Positive (**P**) and negative (**N**) loadings are tested separately. For one-sided components (cell score range ratio >5) the minor side is omitted. Red horizontal line: p=0.05 after Bonferroni correction for multiple testing.

DOI: https://doi.org/10.7554/eLife.43966.009

The following figure supplements are available for figure 3:

Figure supplement 1. Overview of cell score loadings in t-SNE space for all components produced by SDA except single cell components (1, 4, 8, 14, and 46).

DOI: https://doi.org/10.7554/eLife.43966.010

Figure 3 continued on next page

Figure 3 continued

Figure supplement 2. Robustness of SDA Results.
DOI: https://doi.org/10.7554/eLife.43966.011
Figure supplement 3. Rotation Matrix.
DOI: https://doi.org/10.7554/eLife.43966.012
Figure supplement 4. Correlation of C31 gene loadings.
DOI: https://doi.org/10.7554/eLife.43966.013
Figure supplement 5. Robustness of t-SNE embedding.
DOI: https://doi.org/10.7554/eLife.43966.014

additional biological insights for the same number of components (Materials and methods; ***Figure 5D,E,F*** & ***Figure 5—figure supplement 1B***). In addition to obviating the need for further clustering and differential expression analyses, an advantage of using matrix factorization for imputation is the much smaller memory footprint required to store the results: on our dataset MAGIC data is 2.9 Gb whereas the SDA matrices are just 18 Mb (12.6 Mb when loadings with PIP <0.5 are set to 0).

Overall, of 50 components, six represent batch effects, five are components with only a single cell, 13 are observed only in somatic cell types, 23 only in germ cells, and three components load on both somatic and germ cells (***Figure 3—figure supplement 1***). Within somatic cell components, we observe components corresponding to Sertoli cells (n = 4), Leydig cells (4), macrophages (1), T lymphocytes (1), telocytes (1), peritubular myoid cells (1) as well as an interesting component that seems expressed in all interstitial cells (1). Among germ cell-specific components, we observe components corresponding to processes active in spermatogonia (5), preleptotene spermatocytes (1), leptotene/zygotene (2), pachytene (5), diplotene (1), and spermiogenesis (7). Thus, we find multiple sub-components within existing recognized meiotic stages, adding considerable resolution relative to bulk-sequencing approaches. For some analyses below, we considered positively and negatively weighted genes within a component separately, in case these represent different modes of regulation, within the same groups of cells. We provide a web application to enable interactive exploration of gene expression and components at http://www.stats.ox.ac.uk/~myers/testisAtlas.html.

Prior to single cell studies such as our own, previous approaches to germ cell transcriptional profiling provided a single, static summary of pachytene expression from bulk sequencing of purified cells (***da Cruz et al., 2016***; ***Soumillon et al., 2013***). Here, we are able to decompose pachytene gene regulation into five components (13, 39, 42, 47, and 48). Although component cell loadings overlap in pseudotime, they differ dramatically in their dynamics (***Figure 4A&B***). For instance, component 13 and 47 loadings fluctuate between positive and negative, while component 42 loading is constantly negative when active (***Figure 3—figure supplement 1***). The genes with strong loadings within expression components do not necessarily associate with a single, coherent functional process, nor even a set of co-translated transcripts. Instead, components 13, 39, 42 and 48 each involve both a substantial number of genes required for meiosis, but also genes required for postmeiotic processes, including sperm tail formation (***Supplementary file 3***).

Components reflect known biology but also highlight sets of genes with mysterious purpose

Five components correspond to processes in spermatogonia. Component 31 represents undifferentiated spermatogonia expressing *Zbtb16* (aka *Plzf*) (***Buaas et al., 2004***) and *Foxo1* (***Goertz et al., 2011***), while component 50 splits these into two subpopulations one expressing, *Gfra1* (***He et al., 2007***) and *Glis3* (***Kang et al., 2016***), and the other *Nanos3* (***Suzuki et al., 2009***), *Lin28a* (***Zheng et al., 2009***) and *Foxf1* (***Figure 5D***). Component seven likely represents $A_{1\text{-}4}$ spermatogonia expressing *Glis2*, *Nanos1*, *Kit,* and *Stra8*. Component two includes *Ctcfl*, *Pou4f1*, and *Esx1* - likely representing intermediate and type B spermatogonia and component 33 is a broader spermatogonial component enriched in genes involved in spermatogonial differentiation (***Supplementary file 3***).

During meiosis an extended prophase I (lasting 14 days in mice) is itself divided into stages: Leptotene, Zygotene, Pachytene, and Diplotene (***Oakberg, 1956***). During prophase homologous chromosomes pair to enable genetic recombination and balanced segregation during meiotic divisions. In the earlier stages homologous chromosomes begin to associate aided by meiosis-specific cohesin

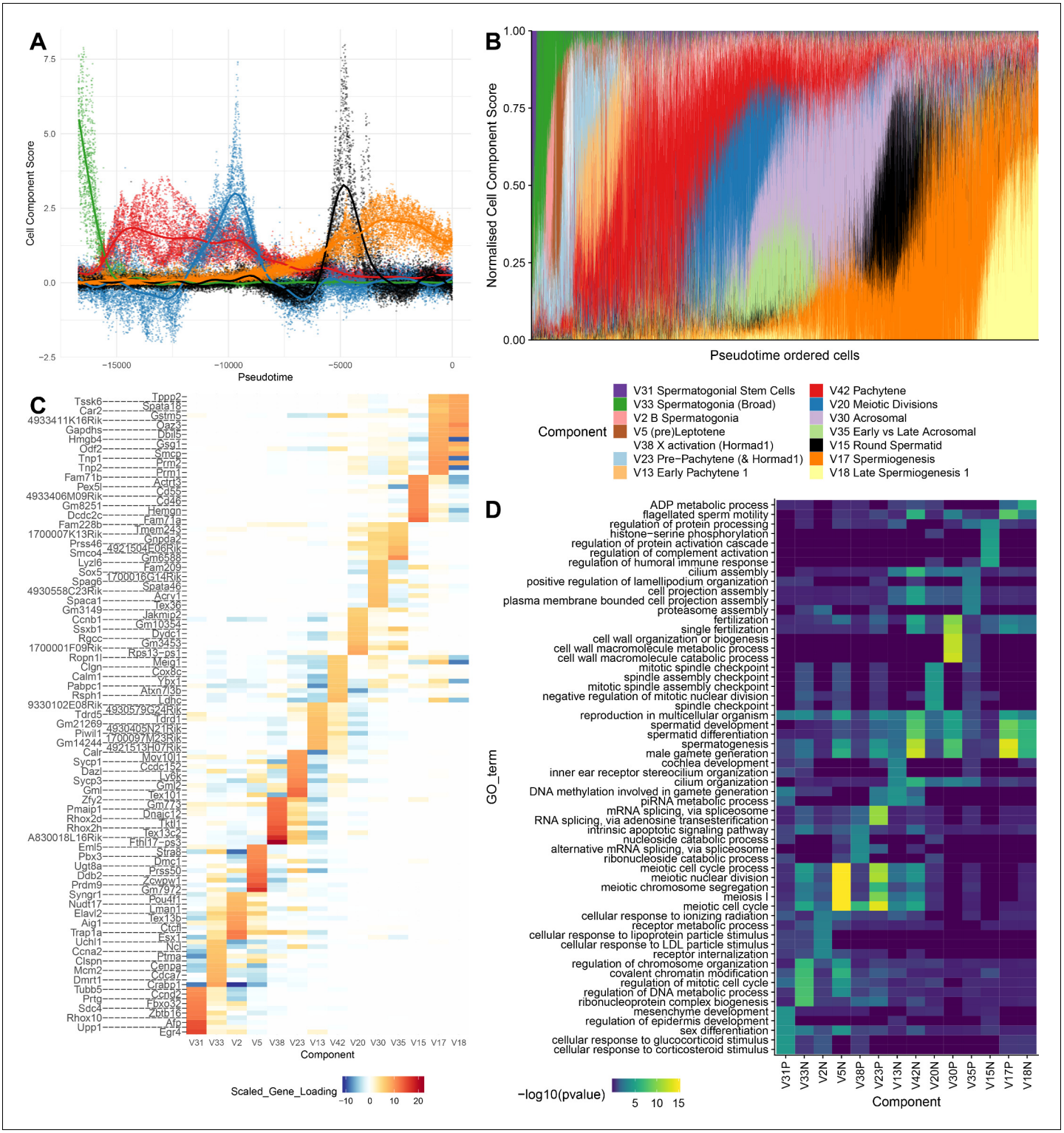


Figure 4. SDA components overlap but represent distinct processes. (**A**) For five example components, the cell scores for each cell are plotted through pseudotime, indicating strongly overlapping dynamically varying component activity. Component signs were chosen to be mainly positive (components have arbitrary sign). Color mappings as in panel B. (**B**) Stacked bar plot of cell component loadings for 14 germ components sorted by cell pseudotime. Each column corresponds to an individual cell and the total positive component loadings for each are normalized to one after flipping components to be mainly positive. Factorization by SDA indicates that transcription during spermatogenesis can be represented as an overlapping series of components in time, coming on and off gradually on different timescales. See also ***Figure 4—figure supplements 1–2*** for alternative visualizations of components in pseudotime. (**C**) Furthermore, these components are comprised of distinct gene sets driving distinct biological processes. Shown are

Figure 4 continued on next page

Figure 4 continued

the top 10 gene loadings for each of the components in (**B**) represented as a heatmap. Most genes have strong loading on only one component. (**D**) Likewise, a gene ontology enrichment analysis for biological processes in the top 250 genes for each component indicates largely non-overlapping enrichments across components. More in-depth analysis of GO enrichments and gene loadings for each component allow separation of components into biological and technical effects (***Figure 4—figure supplements 3–4***).

DOI: https://doi.org/10.7554/eLife.43966.015

The following figure supplements are available for figure 4:

Figure supplement 1. Heatmap of SDA component scores.

DOI: https://doi.org/10.7554/eLife.43966.016

Figure supplement 2. Overview of Individual SDA Components.

DOI: https://doi.org/10.7554/eLife.43966.017

Figure supplement 3. Detailed Analysis of Individual SDA Components.

DOI: https://doi.org/10.7554/eLife.43966.018

Figure supplement 4. Components representing batch effects and cellular respiration.

DOI: https://doi.org/10.7554/eLife.43966.019

and telomeric tethering to the nuclear envelope (***Boateng et al., 2013***; ***Ishiguro et al., 2014***). Several hundred *Spo11*-induced programed double-strand breaks (DSBs) then occour at *Prdm9*-marked sites (***Baudat et al., 2010***; ***Keeney et al., 1997***; ***Myers et al., 2010***; ***Parvanov et al., 2010***). Each DSBs is resected to form single stranded DNA, enabling homology search and repair within the context of a proteinaceous scaffold named the synaptonemal complex (***Zickler and Kleckner, 2015***).

As an illustrative example, we focus on component 5, marking Leptotene. In this component, many of the genes required for these coordinated processes have high (top 500) loadings, including *Prdm9* itself; components of the meiotic cohesin complex *Rad21l*, *Smc1b*, *Smc3*, *Stag3* and *Esco2* (***Rankin, 2015***); components of the telomere tethering complex *Terb1*, *Terb2*, *Spdya,* and *Sun1* (***Ding et al., 2007***; ***Tu et al., 2017***; ***Wang et al., 2019***); genes involved in creating DSBs *Mei1*, *Ccdc36* (*Iho1*), *Spo11* partner *Top6bl* (*Gm960*), and regulator *Atm* (***Lukaszewicz et al., 2018***; ***Reinholdt and Schimenti, 2005***; ***Robert et al., 2016***; ***Stanzione et al., 2016***; ***Vrielynck et al., 2016***); proteins required for the creation and processing of the ssDNA intermediates and their regulators: *Mcm8*, *Dmc1*, *Rad51*, *Rad51ap2, Atr*, *Brca2*, *Tex15*, *Meilb2* (*Hsf2bp*), *Meiob*, and *Spata22* (***Brown et al., 2015***; ***Brown and Bishop, 2015***; ***Dai et al., 2017***; ***Kovalenko et al., 2006***; ***Lee et al., 2015***; ***Martinez et al., 2016***; ***Pacheco et al., 2018***; ***Ribeiro et al., 2018***; ***Widger et al., 2018***; ***Xu et al., 2017***; ***Yang et al., 2008***; ***Zhang et al., 2019***); class I crossover (ZMM group) proteins *Shoc1* (*Zip2* orthologue), *Tex11* (*Zip4* orthologue), *Msh5*, *Hfm1* (*Mer3* orthologue) and regulator *Brip1* (*FancJ*) (***Adelman and Petrini, 2008***; ***Guiraldelli et al., 2018***; ***Guiraldelli et al., 2013***; ***Rakshambikai et al., 2013***; ***Sun et al., 2016***); as well as components of the synaptonemal complex *Sycp1*, *Sycp2*, *Sycp3*, *Syce2*, *Syce3*, *Tex12*, and *Six6os1* (*4930447C04Rik*) (***Gómez-H et al., 2016***; ***Syrjänen et al., 2014***). (***Figure 3D–F***; ***Supplementary file 3***).

Strikingly, this component is highly enriched for GWAS hits of recombination rate in humans (***Halldorsson et al., 2019***). Of the 24 significant GWAS loci identified with confidently associated causal genes, more than half (13) rank within the top 300 genes of this component, and almost all (20) rank within the top 1300 genes ($p=5.2\times10^{-18}$, OR = 77.8 and $p=2.4\times10^{-20}$, OR = 70.1 respectively by Fisher's exact test [FET], ***Figure 3F***). One hit, *Msh4*, is not ranked highly in this component ($2{,}734^{th}$ out of 19,262). However, MSH4 is known to function as a heterodimer with MSH5, ranking 34^{th} (***Rakshambikai et al., 2013***). Unlike GWAS single cell RNA-seq does not rely on the presence of (perhaps rare, small effect) genetic variants for target discovery, while automatically identifying genes rather than SNPs affecting unknown causal genes. For example a previous GWAS (***Kong et al., 2014***) had identified a SNP in the intron of *Ccdc43*, however our expression data strongly suggested the adjacent gene *Meioc* (*aka C17orf104*) as the causal gene (ranked 183^{rd} vs $13{,}651^{st}$ in component 5), providing additional evidence relative to reports that *Meioc* is responsible for maintaining an extended meiotic prophase (***Abby et al., 2016***; ***Soh et al., 2017***). Indeed the lead SNP in this region in a more recent GWAS is in the promoter of *Meioc* (***Halldorsson et al., 2019***). The strong enrichment of genes involved in recombination in this component suggests other highly ranked genes of unknown function could also play key roles in this process. During the

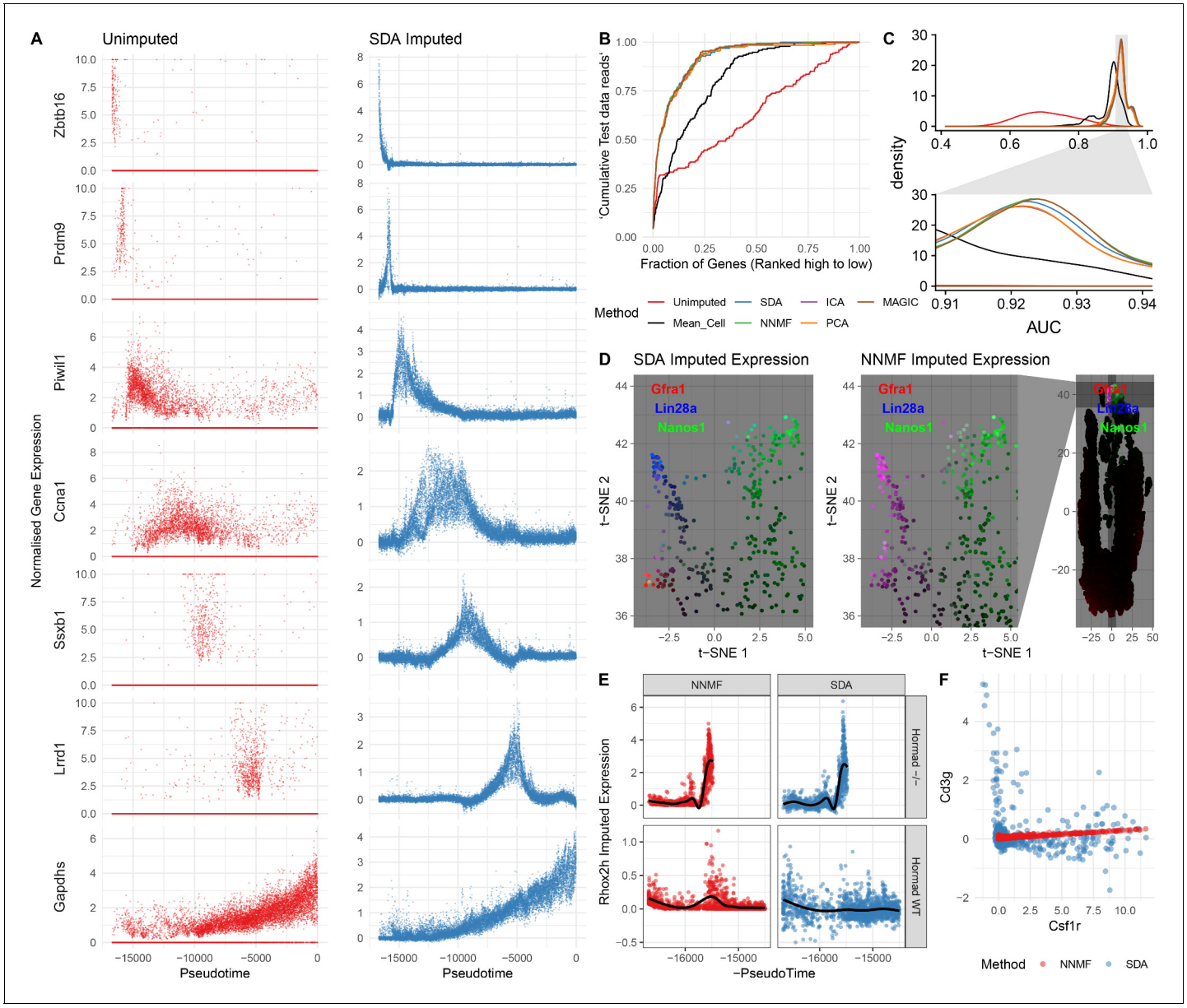


Figure 5. Evaluation of imputation using the SDA model. (**A**) Here, we illustrate the ability of SDA-based imputation (Materials and methods) of gene expression values in single cells to improve the signal/noise ratio of expression, for seven genes with strong developmental regulation. Note in the imputed expression 'dropouts' at 0 are recovered and there is less outlying expression. (**B**) To test the utility of SDA-based imputation, we created separate training/test data (Materials and methods). From the training data we constructed seven predictors of gene expression in the test data for each cell ('Unimputed' using the training data directly, 'Mean Cell' using the mean across all cells, matrix factorisation approaches SDA, PCA, ICA, NNMF, and a dedicated imputation approach, MAGIC). We compared the ability of each predictor to rank the gene expression in the test data for each cell, quantified as the area under the Rank Prediction Accuracy Curve (RPAC). Shown is an example RPAC for these predictors when applied to the test data for a single cell. (**C**) Comparison of AUCs (Area under the RPAC curve) for all cells using various methods (same color scheme as part B). (**D**) SDA produces multiple components for spermatogonia. Shown are zoomed in versions of the t-SNE projection (with full t-SNE for context): cells are colored by expression using a three channel ternary color scheme with the amount of blue, green, red representing the respective expression levels of *Lin28a*, *Nanos1*, and *Gfra1*. By assigning only one component for undifferentiated spermatogonia, NMF predicts *Gfra1* and *Lin28a* are expressed in the same cells resulting in a pink hue (See also ***Figure 5—figure supplement 1B***, no correlation for SDA component 50 Gfra1 Stem Cells). For selection of component see Materials and methods. (**E**) Imputed expression of X chromosomal gene *Rhox2h* from either the SDA or NNMF decomposition, split into cells we know to be either WT or Hormad$^{-/-}$ genotype. NNMF predicts a peak in *Rhox2h* expression even in the WT cells, in which X chromosome activation due to *Hormad1* KO does not occur. (**F**) NNMF does not assign separate components for the innate and adaptive immune cells (See also ***Figure 5—figure supplement 1B***, no correlation for the SDA component 3 Lymphocytes). NNMF does not predict high expression of the adaptive immune cell marker *Cd3g* (T-cell surface glycoprotein CD3 gamma chain), and when it predicts any expression it increases linearly with the innate

Figure 5 continued on next page

Figure 5 continued

immune cell marker *Csf1r* (Macrophage Colony-Stimulating Factor 1 Receptor, or *Cd115*). SDA on the other hand correctly predicts that *Cd3g* and *Csf1r* are not coexpressed in the same cells.

DOI: https://doi.org/10.7554/eLife.43966.020

The following figure supplement is available for figure 5:

Figure supplement 1. Imputation from SDA and Other Matrix Factorization Methods.

DOI: https://doi.org/10.7554/eLife.43966.021

preparation of this manuscript, two such genes were identified: *Ankrd31* (ranked 102nd) plays a role in controlling the number, timing, and location of double strand breaks in meiosis (***Boekhout et al., 2019***; ***Papanikos et al., 2018***), while *Hsf2bp* (now *Meilb2*, ranked 194th) was found to be a master regulator of meiotic recombinases (***Zhang et al., 2019***).

One striking candidate gene is *Zcwpw1*, which ranks 3rd, after *Prdm9*. This gene does not have a known function, but contains two protein domains: CW and PWWP, known to bind H3K4me3 and H3K36me3 respectively (***He et al., 2010***; ***Rona et al., 2016***). PRDM9 deposits both H3K4me3 and H3K36me3 at sites it binds (***Powers et al., 2016***), and this methyltransferase activity is essential for its role in double strand break positioning (***Diagouraga et al., 2018***), suggesting these marks may be recognized by downstream protein(s). An obvious hypothesis is that ZCWPW1 might co-localize to recombination hotspots, by binding the histone modifications deposited by PRDM9. Further work will be required to test this, and the potential role of ZCWPW1 in meiotic recombination.

The early pachytene components 13 and 47 are enriched for genes involved in the meiotic cell cycle (e.g. *Ccna1*, *Cdk1*), chromosome pairing and segregation (e.g. *Sycp3*, *Dmc1*, *Hormad1*), nuclear division (e.g. *Cenpe*, *Plk1*), and piRNA processing (e.g. *Tdrd1*, *Tdrd5, Tdrd9, Piwil1* and *Piwil2*). The next component in the temporal sequence, 48, is restricted to a small cluster of cells in t-SNE space, and enriched for many genes involved in axoneme/cilia assembly (multiple members of the *Cfap* family and dynein genes) and a smaller number of genes involved in microtubule/spindle formation (e.g. *Dcdc2b*, *Ccdc88a*, *Knl1*) and RNA splicing (e.g. *Srrm2*, *Tra2a*, *Srek1*). Components 42 and 39 (pachytene/late pachytene) are enriched for distinct genes, enriched for similar biological functions - such as meiotic cell cycle, cilium assembly, piRNA processing, and translational suppression. These two components, as well as component 47, are significantly enriched for genes that are targets of the transcription factor MYBL1 (as determined by ChIP-Seq, ***Figure 7—figure supplement 1***).

The pachytene components have a striking lack of genes loading on the X or Y chromosome (***Figure 6E***), due to meiotic sex chromosome inactivation (MSCI), which is part of a broader mechanism silencing unsynapsed chromatin (MSUC) (***Turner, 2015***; ***Turner, 2007***). MSCI is an evolutionarily conserved phenomenon essential for proper spermatogenesis in mammals. As previously reported (***Chen et al., 2018***; ***Green et al., 2018***; ***Lukassen et al., 2018***) we observe MSCI from the start of pachytene (***Figure 6A*** and ***Figure 4—figure supplement 3D***). Although previous bulk RNA-seq studies suggested that some genes escape MSCI (***da Cruz et al., 2016***; ***Soumillon et al., 2013***), we were unable to confidently identify any genes escaping MSCI. A small number of sex-chromosome transcripts identified in pachytene cells were observed; however, these genes were highly expressed in neighboring Sertoli cells, suggesting low-level contamination as the most likely explanation. Moreover, our data indicate that previously identified 'escapees' are actually expressed after MSCI, yet fully silenced within MSCI (***Figure 6C*** and ***Figure 6—figure supplement 1B***).

In addition to MSCI there is the potential for lack of sex chromosome transcripts later in post-meiotic cells as they have haploid genomes possessing either an X or a Y chromosome but not both. However, cytokinesis does not fully complete in spermatogenesis resulting in synchronized chains of hundreds of cells, connected by µm-wide cytoplasmic bridges through which mRNA (or perhaps even mitochondria) could be shared (***Greenbaum et al., 2011***). The extent to which mRNA sharing occurs is unknown, but it is a property of interest to evolutionary biology as most models predict a strong fitness benefit to fathers who can mask haploid selection in their gametes (***Otto et al., 2015***). Here, we find that, with respect to sex chromosome transcription, the genetically haploid cells are predominantly phenotypically diploid (***Figure 6A & B***, and ***Figure 6—figure supplement 1A***) suggesting that cytoplasmic mRNA is efficiently shared, consistent with studies of individual genes

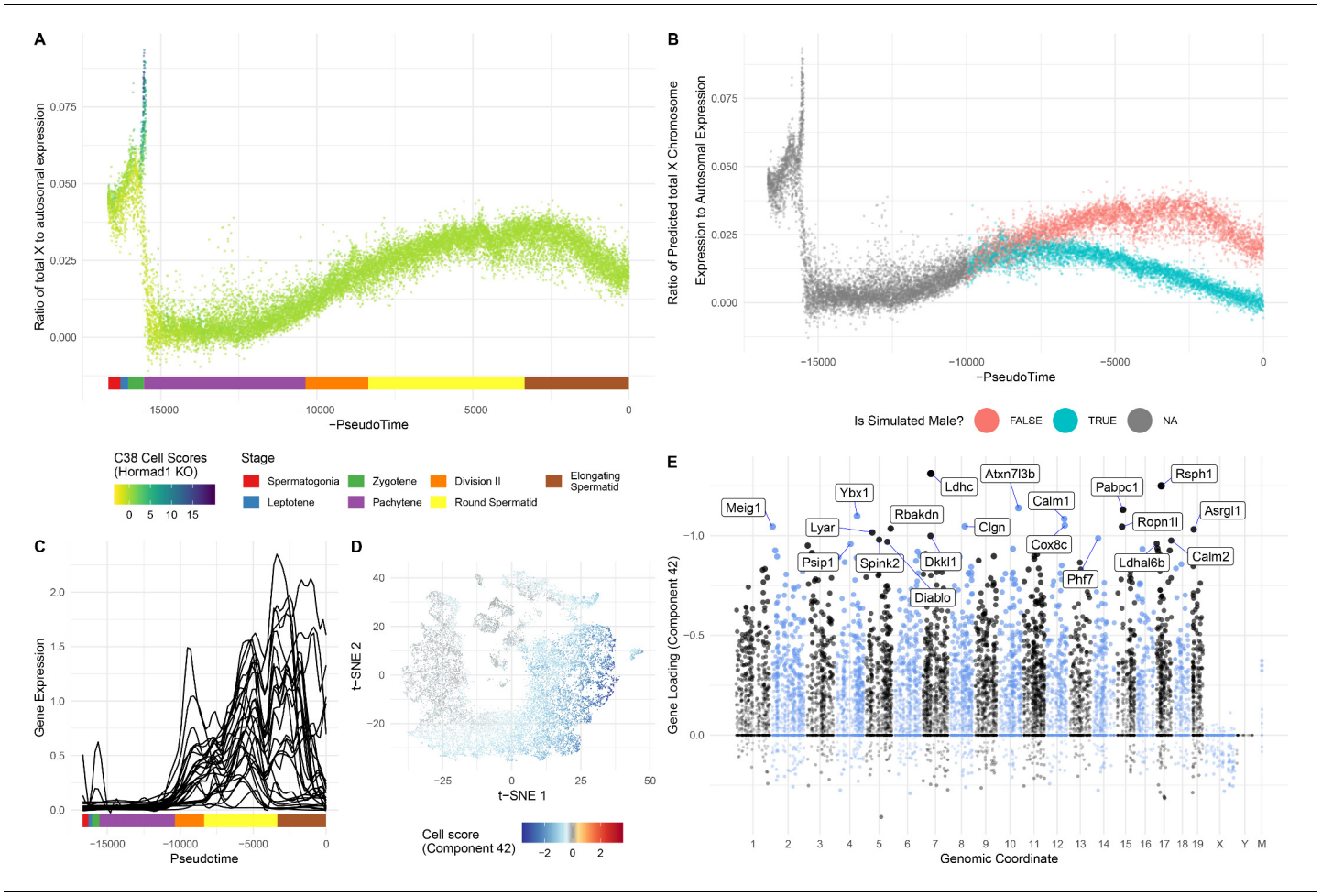


Figure 6. Insights into sex chromosome biology from SDA. (**A**) Pseudotime analysis provides quantitative, high-resolution insights into meiotic sex chromosome inactivation (MSCI). The sum of imputed expression for all genes on the X chromosome divided by that of the autosomes (y-axis) drops to almost 0, showing near-complete MSCI before gradually partially recovering. A similar profile is observed for genes on the Y chromosome (***Figure 6—figure supplement 1A***). (**B**) We do not observe that haploid cells obviously split into two populations due to lack of sex chromosome transcript sharing, in part A. Here were simulate what we might expect to see if there was indeed a lack of sharing (Materials and methods). (**C**) No evidence supporting prior report of genes escaping MSCI. Smoothed expression values (unimputed, gam smoothing with formula 'y ~ s(x, bs = ad)") are shown for each gene reported to escape MSCI (***da Cruz et al., 2016***) excepting *H2al1e*, *H2al1c*, and *Gm10096* which were below our dataset's expression detection threshold. Expression profiles for individual genes are separated in ***Figure 6—figure supplement 1B***. (**D**) Component 42 (Pachytene) cell scores in t-SNE space. (**E**) Component 42 gene loadings. This component represents genes active during the pachytene stage of meiosis; note the striking lack of sex chromosome gene loadings, due to MSCI.

DOI: https://doi.org/10.7554/eLife.43966.022

The following figure supplement is available for figure 6:

Figure supplement 1. Single-gene analysis of MSCI.

DOI: https://doi.org/10.7554/eLife.43966.023

(***Braun et al., 1989***) and recent scRNA-seq reports (***Chen et al., 2018***; ***Green et al., 2018***). However, there remains a possibility that some genes are not shared, such as has been observed for autosomal genes in a mutant heterozygous context: the t-complex responder mutant ($Smok^{Tcr}$) which functions as an antidote in the poison-antidote meiotic drive system of the t-complex (***Véron et al., 2009***) and Spam1 which causes transmission ratio distortion in Robertsonian (Rb) translocation-bearing mice (***Martin-DeLeon et al., 2005***).

Component 20 is particularly interesting, containing genes likely to be functional during meiotic divisions and perhaps afterwards. It contains a number of genes known to be expressed in diplotene and/or key regulators of cell division, in addition to the *Ssx* family of genes (discussed further below)

and also shows very strong enrichment of genes characterized by the presence of a DUF622 domain (18 in the top 88 genes) (**Supplementary file 3**). This rodent-specific gene family arose from duplication of the gene *Dlg5* (**Church et al., 2009**). It was previously shown that many, autosomal, DUF622 genes experience similar epigenetic changes to the sex chromosomes during spermatogenesis (**Moretti et al., 2016**). Another component (9) is most active at a similar time to 20, and is very highly enriched for genes of the electron transport chain ($p=7.4\times10^{-53}$, OR = 104, FET) (**Figure 4—figure supplement 4E&F**).

We identified seven post-meiotic components characterizing wild-type biology. Round spermatid component 30 contains many genes associated with the acrosome, an organelle which forms a nuclear cap containing hydrolytic enzymes used in fertilization (**Ito and Toshimori, 2016**) (**Supplementary file 3**). For one high loading gene, *Lrrc34*, we verified by immunofluorescence that the protein is indeed localized to the acrosome of round spermatids (**Figure 2B**). Component 35, which is essentially concurrent to component 30 in pseudotime, is the most mysterious of all components that we detected. Dozens of protein-coding genes in this component are highly enriched in testis expression but have no known function (**Supplementary file 3**). This component also harbors a substantial number of genes with no apparent ortholog in humans. The existence of such a set of poorly characterized genes likely reflects the difficulty of studying postmeiotic male germ cells - which cannot be differentiated in vitro, host numerous cell-type specific processes, and express many rapidly evolving genes.

The spermiogenesis components 17, 18 and 34 all contain many genes known to be expressed at the latest stages of spermatogenesis, before transcriptional arrest due to replacement of histones with protamines (**Sassone-Corsi, 2002**) (**Supplementary file 3**). In addition, *Abhd5* (aka CGI-58), a protein previously detected in testis lipid droplets (**Wang et al., 2015**), has high loadings specifically in these late components (17 and 18) and we show by immunofluorescence that it serves as an excellent marker of the residual body (**Figure 2B**).

In addition to components for the germ cell transcriptional programs we identified components for at least five different somatic cell types: Sertoli, Leydig, Macrophages, Peritubular Myoid Cells, and T-lymphocytes. We also find a component representing an abundant somatic cell type expressing *Tcf21* but not *Acta2*, described by **Green et al. (2018)** as an unknown mesenchymal celltype, which we identify as telocytes based on coexpression of *Cd34* and *Pdgfra* (**Kuroda et al., 2004**; **Marini et al., 2018**). Some components clearly mark multiple cell types that resolve separately in t-SNE space, while others mark groups of cells that may contain cryptic heterogeneity obscured by overlapping gene expression patterns (**Figure 4—figure supplement 3F** and **Figure 3—figure supplement 1**). We were also able to infer components for batch effects such as differences in sequencing machines and different individual mice (**Figure 4—figure supplement 4A–D**).

Validation and interrogation of SDA components by de novo inference of transcription factor binding sites and comparison to ChIP-seq data

We hypothesized that many of the SDA components represent dynamic and finely tuned transcriptional programs. If this is true, then genes within each component would be expected to have an excess of shared transcription factor binding sites within their cis-regulatory regions. We used an existing approach (**Altemose et al., 2017**; **Davies et al., 2016**) to discover de novo motifs enriched in the promoter regions of the top 250 positive and negative genes (separately) for each component (Materials and methods, **Supplementary file 5**). We compared the resulting motifs with known motifs from the HOCOMOCO database, resulting in 16 groups of matched motifs, including one group of identified de novo motifs not clearly associated with any known transcription factor, but most similar to the binding target of ATF1 (**Figure 7A**, **Figure 7—figure supplement 2**).

Although identified independently, the identified motifs include the binding targets of multiple master regulators of spermatogenesis: *Stra8* (**Kojima et al., 2019**), *Mybl1* (**Bolcun-Filas et al., 2011**), *Rfx2* (**Kistler et al., 2015**), and *Crem* (**Nantel and Sassone-Corsi, 1996**). In addition, we identified the target of *Spi1* (aka *PU.1*), a known master regulator of macrophage differentiation (**Rosa et al., 2007**), specifically in the macrophage component 11. High and specific enrichment of ChIP-seq targets of STRA8, MYBL1, RFX2 and CREM validated our interpretation - that covariation of expression of genes within many components reflects shared transcriptional regulation (**Figure 7—figure supplement 1**).

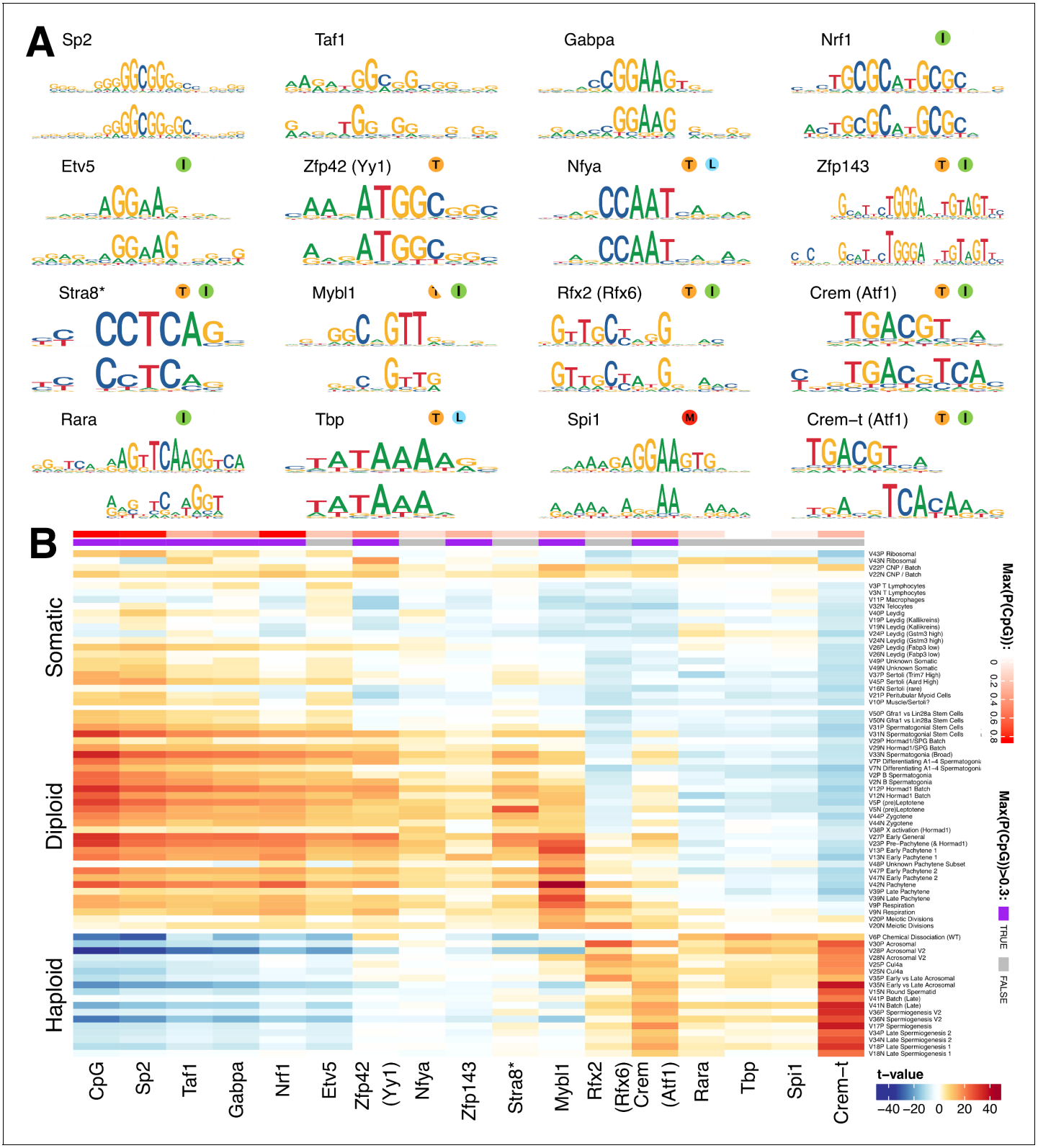


Figure 7. Components show shared *cis*-regulatory features. (**A**) Motifs discovered from the promoter sequences of genes with high component loadings. In each motif logo pair the lower logo shows the de novo inferred motif and the upper logo shows the motif in the HOCOMOCO database best matching the de novo motif. Orange 'T' indicates this transcription factor is highest expressed in testis in the GTEx database (half T indicates second highest). Green 'I' indicates that a mouse knockout of this gene is infertile. Blue 'L' indicates a mouse knockout of this gene is embryonic lethal. Red 'M' indicates this gene is required for macrophage development. The notation 'Crem-t (Atf1)' indicates that we suspect that the true transcription

Figure 7 continued on next page

Figure 7 continued

factor recognizing the motif is not the closest matching database-motif (Atf1). * the (upper) STRA8 motif shown is from Kojima et al., rather than the HOCOMOCO database (**B**) Association of gene loadings with the probability each de novo identified motif is found in the genes for each component. Coloring is a Z-score from a correlation test between gene loadings and motif probabilities,where red (blue) indicates positive (negative) association. The germ cell components (rows) are ordered by pseudotime. The correlation was calculated for positive and negative parts of the component separately and in the cases where the component is mainly one-sided the other side has been omitted, as have the single cell components. The additional column 'CpG' shows the same association test, but with count of promoter CpG dinucleotides, for each component. Across the top of the panel, color bars indicate the maximum probability of there being a CpG at any one position in the denovo motif, and whether that probability is greater than 0.3. See ***Figure 7—figure supplement 2*** for an analogous plot using the HOCOMOCO motif probabilities. We find high and specific enrichment of ChIP-seq targets of STRA8, MYBL1, RFX2 and CREM in the gene loadings of components associated with those motifs, validating our interpretation that covariation of expression of genes within many components reflects shared transcriptional regulation (***Figure 7—figure supplement 1***).

DOI: https://doi.org/10.7554/eLife.43966.024

The following figure supplements are available for figure 7:

Figure supplement 1. Validation of Motif Inference Using ChIP-seq Data.

DOI: https://doi.org/10.7554/eLife.43966.026

Figure supplement 2. Validation of Motif Inference from SDA Component Loadings.

DOI: https://doi.org/10.7554/eLife.43966.025

In our initial analysis, we frequently identified the same motif (e.g. Sp2) in multiple components. Therefore, for each motif-component combination we calculated the association between motif presence at the promoters of genes and the gene loadings (***Figure 7B***). In addition to some relatively specific enrichment (e.g. *Stra8* in Leptotene component 5, *Mybl1* in Pachytene component 42, and *Rfx2* in Acrosomal component 30), this revealed an obvious 'switch' with one group of transcription factors appearing to regulate early meiosis (prior to the meiotic division), and another group regulating meiosis post-division, with only the database Crem and Rfx2 motifs strongly spanning this divide. Moreover, most meiotic motifs spanned several components. This implies that promoter motifs might offer 'broad-scale' control, but differences at 'fine scales' among individual components might frequently also be driven by transcription factor binding to more distant enhancer regions, mRNA degradation by microRNA, or other post-translational mechanisms. Hence, additional work will be required to fully delineate the mechanisms controlling meiotic transcription.

Strikingly, many of the pre-division motifs as well as MLX/CREM contain CpG dinucleotides (***Figure 7***), with most, including the non-CpG exceptions (NFYA and ETV5) also being sensitive to DNA-methylation in their binding (***Domcke et al., 2015***; ***Wang et al., 2017***). None of the post-division motifs contain CpGs, excepting one ATF like motif that we believe is instead Crem-t (see below). Indeed, we found an even stronger pattern of association simply using the count of CpG occurrences as a pseudo-motif (***Figure 7B***), indicating a major shift away from expression of genes whose promoters contain CpG islands, following the meiotic divisions. To find potential effectors of this switch we looked in the component most active at the stage of meiotic division, component 20. We found an enriched family of testis-specific proteins specifically expressed at this time, and characterized by the presence of both the SSXRD and KRAB-related domains. The SSXRD domain has been studied in the context of synovial sarcomas where it was found to associate with the CpG binding protein CXXC2 (KDM2B) which is a component of a non-canonical polycomb complex (***Banito et al., 2018***). Interestingly another non-canonical polycomb component, *Dcaf7*, also has a high loading in component 20 (***Hauri et al., 2016***). It has previously been observed that H3K27me3, a mark deposited by polycomb complex 2, increases dramatically between pachytene and the round spermatid stage (***Sin et al., 2015***). The KRAB-related domain has been studied as part of PRDM9 (the only other gene outside of the X-chromosome cluster to contain both the KRAB-related and the SSXRD domains), where it has been shown to interact with a number of proteins including the CpG binding CXXC1 (***Imai et al., 2017***; ***Parvanov et al., 2017***).

As we inferred these motifs de novo we were able to discover previously unknown motifs. Indeed, we identified a motif in the late components, with partial similarity to the ATF1/CREM motifs but containing an additional CAA tail while mainly lacking the central CpG dinucleotide (***Figure 7A***). Speculatively, this may represent the binding motif of the tau isoform of CREM known to be active

in late spermatogenesis (**Sassone-Corsi, 2000**). Consistent with the more general pattern of CpG occurrence we find this ATF1/CREM-t motif highly associated with post-division components: in fact it is the most strongly associated motif in multiple such components (**Figure 7B**).

Joint analysis of 5 mouse strains identifies pathology-related components

The flexibility of the SDA modeling framework allows the identification of sets of genes that show significant covariation in small numbers of cells. Thus, a joint analysis of mutant and wild-type cells using SDA could potentially decompose expression variation into separate technical effects, variation due to normal biological processes, and variation due to pathology, identifying mutant-specific components in the context of wild-type cells. We set out to evaluate the utility of single-cell sequencing to identify pathology in each mutant strain, combining results from both classical and SDA approaches.

Increased apoptosis is an important mechanism underlying many genetic forms of male infertility in mice. Apoptotic cells can be identified from single-cell RNA-seq data as having an excessive proportion of total transcriptome derived from mitochondrial genes (**Ilicic et al., 2016**). Cells from *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ animals showed higher rates of apoptosis compared to wild-type, *Cul4a* and *Cnp* (2% vs 14.5%, **Figure 1—figure supplement 1**). Pseudotime analysis provided an even finer level of resolution for staging the time of onset of developmental problems in each strain (**Figure 8A**). By performing joint pseudotime analysis on all strains simultaneously, it is in theory possible to fine map the timing of developmental defects. Our our pseudotime-ordered set of 16,950 germ cells spans the entire ~34.5 day (**Oakberg, 1957**) development process from Type A spermatogonia to mature spermatozoa, suggesting a mean difference in developmental age between pseudotime-adjacent cells of 3 minutes. Although further work is needed to clarify the mapping of pseudotime to real time, that mapping estimates the difference in the mean time of arrest of *Hormad1*$^{-/-}$ cells and *Mlh3*$^{-/-}$ cells to be 12 days. This difference is reflected in the SDA components as well; *Mlh3*$^{-/-}$ animals possess cells that load on pachytene components 47, 42 and 39, while *Hormad1*$^{-/-}$ animals do not.

HORMAD1 is a meiosis-specific protein that regulates chromosome recombination, synapsis, and segregation. HORMAD1 normally marks un-synapsed chromosomes (including sex chromosomes). While HORMAD1 is removed by TRIP13 on synapsis, it persists on asynapsed chromosomes, which then undergo MSUC, leading to MSCI for the sex chromosomes (**Shin et al., 2010**; **Wojtasz et al., 2009**). In *Hormad1*$^{-/-}$ spermatocytes, double-strand break formation and early recombination are disrupted as marked by the reduction of γH2AX, DMC1, and RAD51 foci (**Shin et al., 2010**). Hard clustering analysis (**Figure 1F & G**) showed a deficit of post-pachytene *Hormad1*$^{-/-}$ germ cells, consistent with the expectation that *Hormad1*$^{-/-}$ mutant cells experience apoptosis during meiosis I due to pachytene checkpoint failure (**Daniel et al., 2011**). Along with this arrest phenotype, the *Hormad1*$^{-/-}$ leptotene/zygotene cells form a distinct cluster outside of the leptotene/zygotene cells of all other strains (Cluster 30, **Figure 1—figure supplements 2** and **3**). A list of significant differentially expressed genes between the cluster 30 and neighboring cluster 32 included a number of sex chromosome genes (**Supplementary file 2**). Consistent with these observations, we found one SDA component (38) with much higher gene loadings on the sex chromosomes than autosomes (**Figure 9A**, **Figure 4—figure supplement 3C**, **Supplementary file 3**), and with cell loadings that are specific to *Hormad1*$^{-/-}$. We find that not only does *Hormad1*$^{-/-}$ fail to silence previously expressed sex-linked genes, but many previously *un*expressed sex-linked genes such as *Rhox2h* obtain high expression (**Figure 9B**). Interestingly, there are also multiple autosomal genes with high loadings. This may be due to ectopic expression of sex-linked transcription factors; for example, *Zfy1* and *Zfy2* were previously shown to cause pachytene arrest when misexpressed (**Royo et al., 2010**). We find a very strong association between genes in this component and genes overexpressed in mice which have mutations in either *Hormad1* or *Trip13* ($p = 2.2\times10^{-39}$, OR = 184 and $p = 1.3\times10^{-157}$, OR = 115, respectively by FET) (**Ortega, 2016**; **Figure 4—figure supplement 3A&B**).

CUL4A is a major component of the E3 ubiquitin ligase complex called CRL4 which is known to regulate cell cycle, DNA replication, DNA repair, and chromatin remodeling (**Dubiel et al., 2018**). Studies on *Cul4a*$^{-/-}$ mice noted that some spermatocytes arrest at the pachytene stage of meiosis I induced by the pachytene checkpoint, whereas remaining spermatocytes complete meiosis but the resulting spermatozoa present oligoasthenospermia and severe malformations (**Yin et al., 2011**).

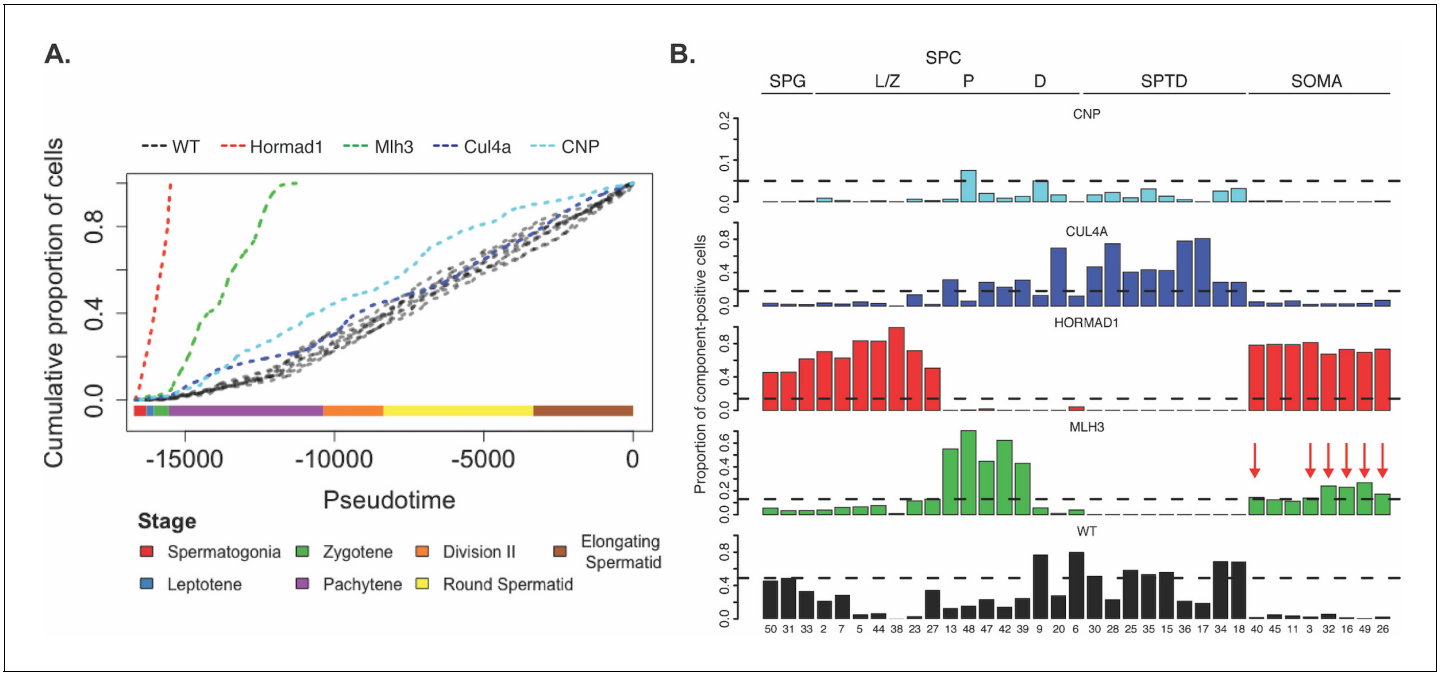
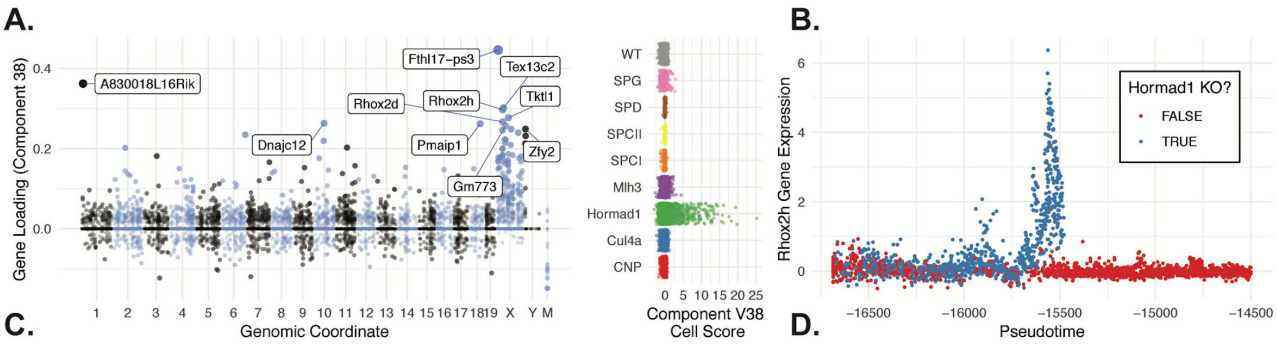
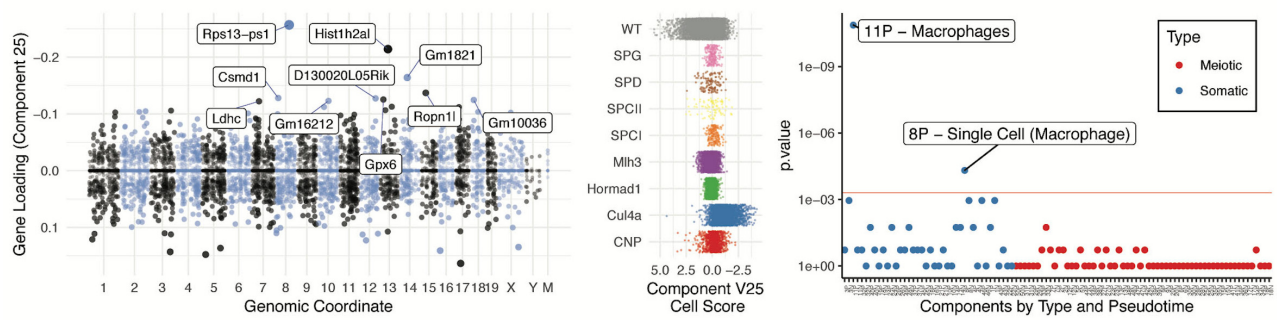


Figure 8. Characterization of mouse mutants with testicular phenotypes using pseudotime and SDA. (**A**) The cumulative distribution of cells along pseudotime from each mouse strain. The data clearly indicate that *Hormad1*$^{-/-}$ cells arrest prior to *Mlh3*$^{-/-}$ cells in the pachytene stage of spermatogenesis, while *Cul4a*$^{-/-}$ and *CNP* mice show quantitative deviation from WT in the abundance of postmeiotic cells. (**B**) As a way to summarize the SDA analysis of each strain, we plot the proportion of cells with strong component loadings from each strain separately. If cells are randomly distributed across components then we would expect the fraction of cells from each mutant to be the proportion of total cells sequenced from that mutant (dashed horizontal lines). Instead there are clear enrichments of component loadings in particular mutants, providing a fingerprint of pathology for those strains. SDA components are sorted by developmental stage, as indicated by horizontal lines across the top of the panel.
SPG = spermatogonial components; L/Z = leptotene/zygotene components; P = pachytene components; D = diplotene components; SPTD = components in spermiogenesis; SOMA = somatic cell components.
DOI: https://doi.org/10.7554/eLife.43966.027

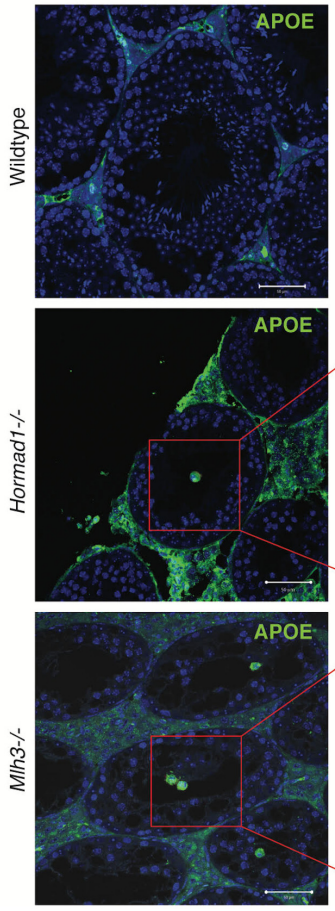
The molecular basis of observed abnormal phenotypes in spermatozoa remains unclear. We identified a single SDA component (25) that was highly specific to *Cul4a*$^{-/-}$ cells (***Figure 9C*** and ***Supplementary file 3***). This component corresponds to dozens of genes that are overexpressed in *Cul4a*$^{-/-}$ mutants when compared to all other strains, with GO enrichments related to spermatid development, motility and capacitation. These findings are consistent with the observed phenotype of *Cul4a*$^{-/-}$ mice and provide new leads to investigate mechanisms of pathology.

MLH3 is an essential protein required for crossover formation in early meiosis and for binding of MLH1 to meiotic chromosomes. Studies on *Mlh3*$^{-/-}$ testes have shown depletion of spermatocytes and some spermatogonia due to apoptosis in diplonema induced by a reduction of chiasmata and a loss of recombination nodules (***Lipkin et al., 2002***). Interestingly, in contrast to *Hormad1*$^{-/-}$, we found no obvious transcriptional phenotype in *Mlh3*$^{-/-}$ cells either by SDA analysis or by comparison of expression levels between hard-clustered wild-type and mutant cells (other than differential expression of *Mlh3*). Instead, *Mlh3*$^{-/-}$ spermatocytes might simply trigger apoptosis through existing checkpoint protein machinery assembled earlier in development. Using the simple pseudotime analysis described above, we can estimate that if a transcriptional response was triggered, it might last less than approximately half an hour, for it to be missed in our sample of cells (***Figure 8A***). Similarly, the cells from *Cnp* mutant mice did not form distinct clusters, nor did they show SDA component loadings distinct from wild-type cells. Although the presence of multinucleated giant cells, hypocellular seminiferous tubules and infertile phenotype in these mice point to a serious defect in spermatogenesis, it seems difficult to determine which stages are affected using single-cell expression data. One possible explanation of missing important biological signals may be that *Cnp* mice present a partial arrest phenotype which masks the developmental abnormalities. Another possible





E.



	APOE positive tubules	Total tubules	Percent	P-value vs WT
Wildtype	0	2496	0	-
Hormad1-/-	25	707	3.5%	<2.2x10-16
Mlh3-/-	98	1756	5.6%	<2.2x10-16

Figure 9. Dissection of strain-specific pathology. (**A**) SDA component 38 is comprised largely of genes on the X chromosome, with a gene loading direction that indicates failure of X inactivation. As illustrated by the cell scores (loadings) for this component, it is restricted to *Hormad1*$^{-/-}$ cells. (**B**) Pseudotime analysis indicates that *Hormad1*$^{-/-}$ cells diverge developmentally from all other strains around leptotene/zygotene. In this illustration, the X-linked gene *Rhox2h* is shown to have low or no expression in all cells prior to meiosis, and then rapidly increased expression specifically in *Hormad1*$^{-/-}$ cells until this lineage arrests. (**C**) Component 25 is the component most strongly enriched for *Cul4a*$^{-/-}$ cells. (**D**) We identified six components with shared enrichment for both *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ cells; these components contained genes with numerous significant GO associations related to Alzheimer's disease (AD) pathology (main text, ***Figure 8B***). For each SDA component, we tested for association between known AD genes and genes with either positive (P) or negative (N) loadings on that component. AD genes are highly enriched for expression in component 11, corresponding to macrophages. (**E**) Further investigation of protein expression of AD genes revealed APOE+ (green) cells within the tubules of *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ but not WT. These cells showed nuclear morphology different from native germ cells or Sertoli cells, and stain positive for the macrophage marker F4/80. The inset table summarizes raw data on the frequency of APOE+ tubules obtained by microscopy. The frequency of APOE+ tubules is more common in each mutant strain when compared to WT by Fisher's exact test. Scale bar = 50 μm.

DOI: https://doi.org/10.7554/eLife.43966.028

explanation is that droplet-based sequencing library preparation may undersample the cells with aberrant transcriptional signatures, for example due to failure of oil droplets to encapsulate the giant cells.

Invasion of macrophages into the seminiferous tubules is a convergent phenotype of meiotic arrest mutants

Despite the differences in cell composition or component loadings among mutant strains, we identified six somatic components (3, 16, 49, 40, 26, and 32) showing a specific enrichment for *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ cell loadings when compared to all other strains (***Figure 8B***). Hypothesis-free GO enrichment analysis of these components (Materials and methods) revealed a recurrence of amyloid related GO terms with qvalue <0.01, with these terms being the highest enriched term in three components (26N, 49N, 16N, ***Supplementary file 6***). Excessive production of amyloid-beta, a primary cause of Alzheimer's disease, was not previously reported in these mutants, and the possible physiological role of such production is unclear. We tested multiple antibodies to human amyloid-beta that failed to work on our tissue. To further evaluate the expression of Alzheimer's disease (AD)-related genes across all five mouse strains, we tested individual SDA components for enrichment of expression of AD risk genes identified in a recent GWAS, identifying component 11 (macrophages) as specifically and strongly enriched ($p=1.3x10^{-11}$, OR=65.9 by FET, ***Figure 9D*** and ***Figure 4—figure supplement 3E***). Immunofluorescence staining for the protein product of one well-studied AD gene, *Apoe*, in wild-type animals showed low levels of specific staining, confined to the interstitial space (***Figure 9E***). Both *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ displayed interstitial cell with more intense staining of APOE, as well as a greater abundance of APOE+ cells. More surprisingly, we also found a rare population of APOE+ cells within the tubules of *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$, that was never observed in wild-type. We screened 4959 tubule cross-sections to establish more precise estimates of APOE+ cell frequency in these three lines (Materials and methods). When compared to the frequency in wild-type tubules (0/2496 tubules), we see higher frequencies of intratubular APOE+ cells in *Mlh3*$^{-/-}$ (25/707 tubules, 3.5%, $p<2.2\times10^{-16}$) and *Hormad1*$^{-/-}$ (98/1756 tubules, 5.6%, $p<2.2\times10^{-16}$). These APOE+-cells displayed a nuclear staining and morphology that are distinct from normal germ cells and Sertoli cells, and appeared more similar to APOE+ cells outside of the tubules. These APOE + intratubular cells stained for F4/80, a well-established macrophage antigen, perhaps surprisingly, given that it suggests that in these mutants, immune cells can transit the blood-testis barrier and enter an area typically regarded as immune-privileged. Intratubular macrophages have rarely been described previously, again nearly always in the context of testicular defects (***Frungieri et al., 2002***; ***Goluža et al., 2014***; ***Holstein, 1978***). Co-staining of F4/80 with an antibody for activated CASPASE-3, a marker of apoptosis, failed to identify any double positive cells, excluding the possibility that intratubular F4/80 protein expression was somehow an artifact of an apoptotic cell population. The mechanisms by which macrophages transit the blood-testis barrier, and the corresponding cues for migration, await further investigation.

Discussion

The extensive cellular heterogeneity of the testis has limited the application of genome technology to the study of its gene regulation and pathology. Here, we described how the SDA analysis framework can be applied to single-cell RNA-sequencing data of the testis to overcome the challenge of heterogeneity by summarizing gene expression variation into components that reflect technical artifacts, cell types, and physiological processes. Rather than clustering groups of *cells*, SDA identifies components comprising groups of *genes* that covary in expression, and represents a single cell transcriptome as a sum of such components. This revealed previously uncharacterized complexity, with multiple different components even within recognised meiotic stages such as pachytene. This finer granularity suggests new biological interactions, for example the extremely high expression of *Zcwpw1*, a reader of specific histone modifications, within the same component as *Prdm9*, which induces identical modifications. We also identified components, both meiotic and non-meiotic, corresponding to interpretable pathology and specific to one or more mutant strains.

Other matrix factorization methods have been previously applied to soft cluster high dimensional gene expression data for example ICA, PCA (**Alter et al., 2000**; **Green et al., 2018**), Bayesian Factor Analysis (**Bernardo, 2003**) and Non-Negative Matrix Factorization (NNMF) (**Brunet et al., 2004**; **Kim and Tidor, 2003**) which naturally has a degree of sparsity in both the cell scores and gene loadings due to the positivity constraint. More recently these methods have also been applied to single cell RNAseq data (**Duren et al., 2018**; **Kotliar et al., 2018**; **Saunders et al., 2018**; **Shao and Höfer, 2017**; **Welch et al., 2019**; **Zhu et al., 2017**, reviewed in **Stein-O'Brien et al., 2018**). Here, we have reported some comparisons between SDA and these standard methods. NNMF is often motivated by the positive nature of the original data, in addition to potentially increased interpretability for purely positively additive components. However, we note that latent factors, such as those utilized by SDA, which allow negative loadings have the potential to better capture transcriptional repression. In our tests, SDA retains similar imputation performance to NNMF, while providing a more compact (in terms of sparsity) representation of the data – aiding our interpretation of components found. Beyond matrix factorization, there are other frameworks with similar goals that have been applied successfully to single cell data. One set of methods are those based on neural networks, such as self-organizing maps (**Löffler-Wirth et al., 2015**) and deep-network autoencoders (DCA) (**Eraslan et al., 2019**). DCA, much like t-SNE, creates a nonlinear embedding of the high dimensional data resulting in a lower dimensional set of scores for each cell. This approach does not, however, provide the equivalent to gene loadings and so one would have to do additional differential expression analysis on a hard clustering of the latent embeddings in order to find genes associated with the latent dimensions. To assist comparison of SDA to other methods with overlapping objectives, we have summarized resource usage of SDA across a variety of run parameters, and input data sizes (**Supplementary file 7**).

By performing de novo motif analysis, we observed that it is possible to identify transcription factors critical for the meiotic program without prior knowledge, as well as other motifs not currently well characterized. It seems very likely that our analysis of promoters is only a first step towards what is possible here via – for example – analysis of enhancers and other regulatory sequences, and we hope that future data will allow this, working towards identifying the full set of transcription factors, and their targets, used in mammalian spermatogenesis. The apparent dramatic change away from the use of factors binding CpG dinucleotides, and whose binding is often disrupted by methylation of such dinucleotides, after the first meiotic division, is one area for such further research – whether this involves *Ssx* genes, DUF622-containing genes, and/or other factors. More generally, in combination with temporal information from pseudotime analysis, it will be possible to create a model of the cascade of gene regulation, and by comparison across species, better understand the constraints on the precise timing and ordering of regulatory events.

Finally, we note that gene expression components (for example those identified by SDA) represent an attractive way to build a dictionary of pathology of the testis. The construction of new component models using a larger panel of mutants with known pathologies will accelerate the interpretation of idiopathic mutants, and, ultimately, could provide a framework for a much more advanced diagnosis of human infertility than is currently in practice.

Materials and methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Genetic reagent (*M. musculus*)	C57BL/6J	Jackson Laboratory	Cat# 000664	
Genetic reagent (*M. musculus*)	B6.129-Mlh3tm1Lpkn/J	Jackson Laboratory	Cat# 018845	***Lipkin et al., 2002***
Genetic reagent (*M. musculus*)	B6;129S7-Hormad1tm1 Rajk/Mmjax	Jackson Laboratory	Cat# 41469-JAX	***Shin et al., 2010***
Genetic reagent (*M. musculus*)	B6;129-Cul4a$^{-/-}$	PMID:21624359		Liang Ma Lab (WUSTL)
Genetic reagent (*M. musculus*)	C57BL/6J CNP eGFP BAC TRAP	this paper		Joseph Dougherty Lab (WUSTL)
Genetic reagent (*M. musculus*)	B6;CBA-Tg(Pou5f1-EGFP)2Mnn/J	Jackson Laboratory	Cat# 004654	
Antibody	Rabbit polyclonal anti-LRRC34 (G-15)	Santa Cruz Biotechnology	Cat# sc-99549, RRID:AB_2137597	(1:100) dilution
Antibody	Mouse monoclonal anti-NOP132/ NOL8 (D-7)	Santa Cruz Biotechnology	Cat# sc-390011	(1:100) dilution
Antibody	Mouse monoclonal anti-UKHC/ KIF5B (F-5)	Santa Cruz Biotechnology	Cat# sc- 133184, RRID:AB_2132389	(1:100) dilution
Antibody	Mouse monoclonal anti-ABHD5 (E-1)	Santa Cruz Biotechnology	Cat# sc- 376931	(1:100) dilution
Antibody	Rat monoclonal anti-F4/80 (BM8)	Santa Cruz Biotechnology	Cat# sc-52664, RRID:AB_629466	(1:100) dilution
Antibody	Rabbit polyclonal anti-cleaved Caspase-3 (Asp175)	Cell Signaling Technology	Cat# 9661	(1:400) dilution
Antibody	Mouse monoclonal anti-ApoE (HJ6.3)	David Holtzman Lab (WUSTL), PMID: 23129750		(1:1000) dilution
Antibody	Goat polyclonal anti-VIMENTIN (C-20)	Santa Cruz Biotechnology	Cat# sc-7557, RRID:AB_793998	(1:100) dilution
Antibody	Goat polyclonal anti-ACYP1 (K-13)	Santa Cruz Biotechnology	Cat# sc-160129, RRID:AB_2242291	(1:100) dilution
Antibody	Goat polyclonal anti-CCDC62 (N-15)	Santa Cruz Biotechnology	Cat# sc-240210	(1:100) dilution
Antibody	Goat polyclonal anti-UNC80 (L-16)	Santa Cruz Biotechnology	Cat# sc-165859	(1:100) dilution
Antibody	CF594 donkey anti-goat	Biotium	Cat# 20116, RRID:AB_10559039	(1:300) dilution

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Antibody	Alexa Fluor 488 donkey anti-mouse	Life Technologies	Cat# A-21202, RRID:AB_141607	(1:300) dilution
Antibody	Alexa Fluor 488 donkey anti-rabbit	Life Technologies	Cat# A-21206, RRID:AB_2535792	(1:300) dilution
Antibody	Alexa Fluor 594 donkey anti-mouse	Life Technologies	Cat# A-21203, RRID:AB_2535789	(1:300) dilution
Antibody	Alexa Fluor 594 donkey anti-rabbit	Life Technologies	Cat# A-21207, RRID:AB_141637	(1:300) dilution
Sequence-based reagent	Drop-seq beads	ChemGenes	Macosko201110	PMID:26000488
Sequence-based reagent	Drop-seq reagents	PMID:26000488		
Commercial assay or kit	NexteraXT	Illumina	Cat# FC-131–1024	
Commercial assay or kit	Medimachine	BD Biosciences	Cat# 340588	
Commercial assay or kit	50 um Medicon	BD Biosciences	Cat# 340591	
Chemical compound, drug	Hoechst 33342	Invitrogen	Cat# H3570	
Software, algorithm	Zen	other	RRID:SCR_013672	https://www.zeiss.com/microscopy/us/products/microscope-software/zen.html; RRID:SCR_013672
Software, algorithm	STAR	PMID:23104886	https://github.com/alexdobin/STAR	**Dobin et al., 2013**
Software, algorithm	Drop-seq_tools	PMID:26000488	http://mccarrolllab.com/dropseq/	**Macosko et al., 2015**
Software, algorithm	Picard Tools	other	http://broadinstitute.github.io/picard/	
Software, algorithm	Samtools	other	http://samtools.sourceforge.net	
Software, algorithm	Seurat	PMID: 29608179	http://satijalab.org/seurat/	**Butler et al., 2018**
Software, algorithm	SDA	PMID: 27479908	https://jmarchini.org/sda/	**Hore et al., 2016**
Software, algorithm	SDAtools	PMID: 27479908	https://github.com/marchinilab/SDAtools (archived at 10.5281/zenodo.3233974)	**Hore et al., 2016**
Software, algorithm	TomTom	PMID: 17324271	http://meme-suite.org/tools/tomtom	**Gupta et al., 2007**
Software, algorithm	MotifFinder	PMID: 29072575	https://github.com/MyersGroup/MotifFinder (archived at http://dx.doi.org/10.5281/zenodo.3234026)	**Altemose et al., 2017**
Software, algorithm	ggseqlogo	PMID: 29036507	https://omarwagih.github.io/ggseqlogo/	**Wagih, 2017**
Software, algorithm	edgeR	PMID: 19910308	https://bioconductor.org/packages/release/bioc/html/edgeR.html	**Robinson et al., 2010**

Mice

All animal experiments were performed in compliance with the regulations of the Animal Studies Committee at Washington University in St. Louis under protocol #20160089. Mice were housed in a barrier facility under standard housing conditions with *ad libitum* access to food and water and a 12 hr:12 hr light/dark cycle. All single-cell RNA sequencing experiments were carried out with sexually mature animals (ages of mice in this paper vary from 11 to 38 weeks) except for *Pou5f1-EGFP* transgenic animal testes which were collected at post-natal age (P) 7. For specific age of mouse at the time of testes collection for different batches, please refer to ***Supplementary file 1***. Samples for histological studies were also collected at the time of testes collection for single-cell RNA sequencing. The mouse lines used in this paper are the following:

1. C57BL/6J male mice were used for Hoechst-FACS and total testis single-cell RNA sequencing experiments.
2. B6;CBA-Tg(*Pou5f1*-EGFP)2Mnn/J reporter mice were used for enriching and isolating spermatogonia type A cells. Testes from five mice at post-natal age P7 were pooled to generate single-cell suspension and FACS sorted for GFP positive cells, followed by Drop-seq.
3. B6.129-*Mlh3*^{tm1Lpkn}/J heterozygotes were bred to maintain the colony and male homozygotes were used for Drop-seq experiments.
4. B6;129S7-*Hormad1*^{tm1Rajk}/Mmjax heterozygotes were bred to maintain the colony and male homozygous knockouts were used for Drop-seq experiments.
5. B6;129 *Cul4a*$^{-/-}$ mice were used for generating Drop-seq data
6. C57BL/6J CNP-EGFP BAC-TRAP mice were used for Drop-seq data

Single-cell suspension preparation

Mechanical dissociation of testes

Two different types of testicular dissociation protocols were used in this paper: enzymatic and mechanical. Both enzymatic and mechanical protocols were previously published in ***Getun et al. (2011)*** and ***Geisinger and Rodríguez-Casuriaga (2010)***. These methods were modified appropriately for single-cell RNA sequencing. For mechanical dissociation method, fresh testes were decapsulated in 1X DMEM and cut into small pieces (approximately 2–3 mm^3). These tissue fragments were transferred to a 50 µm Medicon, a tissue disaggregator and tissue fragments were dissociated in 1 mL 1X DMEM for 5 min on Medimachine. The resulting single-cell suspension was aspirated from Medicon with a 3 mL needless syringe. This dissociation/aspiration step was repeated three times and total of 3 mL single-cell solution was retrieved. Then the single cells were filtered through sterile 40 um strainers twice and triturated for 1 min with a wide orifice disposable Pasteur pipet. Cells were spun down at 500xg for 10 min at 4˚C and re-suspended in 2 mL 1X DMEM. Finally, cells were filtered once more with sterile 50 um filter, adjusted to 100 cells/µl concentration, and placed on ice until processed for Drop-seq. Single-cell RNA sequencing experiments were performed within ~30 min of testes collection for mechanical dissociation.

Enzymatic dissociation of testes

Solutions necessary for enzymatic dissociation were prepared fresh prior to testes collection and these solutions are as follows: 120 U/mL collagenase type I in 1X DMEM; 50 mg/mL trypsin in 1 mM HCl; 1 mg/mL DNase I in 50% glycerol. For enzymatic dissociation method, decapsulated fresh testes were collected in 15 mL conical tubes, one testis per tube. Each testis was dissociated in 6 mL of collagenase type I solution and 10 µl of DNAse I solution with horizontal agitation at 120 rpm for 15 min at 37˚C. Tubules were decanted for 1 min vertically at room temperature and supernatant was discarded. Another 4 mL of collagenase type I solution, 50 µl of trypsin solution and 10 µl of DNAse I solution were added to each tube and incubated with horizontal agitation at 120 rpm for 15 min at 37˚C. Testicular tubules were triturated with a plastic disposable Pasteur pipet with a wide orifice for 3 min. Another 30 µl of Trypsin solution and 150 µl of DNAse I solution were added and incubated for 10 min with horizontal agitation at 120 rpm. Then 400 µl Fetal Bovine Serum (FBS) was added to deactivate dissociation enzymes. Finally, collected single-cell suspension was passed through 40 µm filter twice and stored on ice until processing for Drop-seq. These cells were processed within 1.5 hr of the testes collection.

For digesting *Pou5f1*-EGFP mice testes, we adapted a protocol described previously (**Garcia and Hofmann, 2012**). Briefly, testicular tubules/fragments were incubated in 200 µg/mL trypsin solution for 15–20 min with intermittent pipetting followed by 300 µl FBS addition for inactivating trypsin. Single-cells suspension was filtered through 50 µm filters twice and stored on ice until FACS.

Isolation of germ cell populations by flow cytometry

Hoechst-FACS for spermatocytes and spermatids

For isolation of major germ cell populations, we adapted a Hoechst-FACS protocol and sequential gating strategies described in **Lima et al. (2017)**. Briefly, 10 µl Hoechst and 2 µl of propidium iodide (PI) were added to single-cell suspension obtained from one testis and incubated at room temperature for 20 min. Then single-cell suspension was filtered through a 50 um cell strainer. Cells were sorted and analyzed using Beckman Coulter MoFlo Legacy cell sorter and Summit Cell sorting software. First, debris were excluded based on forward scatter (FSC) and side scatter (SSC) plot pattern. Single cells were gated by adjusting FSC and pulse width threshold. Dead cells were gated and removed based on PI intensity. A minimum of 500,000 events were observed before proceeding to gating on different germ cell populations. Then, cell count histogram was plotted based on Hoechst blue fluorescence and observed three peaks, representing haploid (1C), diploid (2C), and tetraploid (4C) populations. Then Hoechst-blue and Hoechst-red fluorescence intensities were plotted to refine spermatocytes and spermatids populations.

Spermatogonia type A

For isolating spermatogonia type A cells from the *Pou5f1-EGFP* reporter mice, cells were analyzed and sorted with the same cell sorter and software described above section. Similar sequential gating strategies were followed. Debris were excluded, single cells were gated and dead cells were excluded. Then, GFP+ cells were gated on a plot of GFP vs FSC.

Single-cell RNA sequencing library generation

Drop-seq procedure

Drop-seq sequencing libraries were generated according to the protocol described previously (**Macosko et al., 2015**). Cells and beads were diluted to co-encapsulation occupancy of 0.05. Two bead lots were used for generating Drop-seq data (For more details, see **Supplementary file 1**). Individual droplets were broken by perfluorooctanol, followed by bead harvest and reverse transcription of hybridized mRNA. After Exonuclease I treatment, aliquots of 2000 beads were amplified for 14 PCR cycles (all necessary PCR reagents and conditions were identical to **Macosko et al., 2015**). PCR products were purified using 0.6x AMPure XP beads and cDNA from each experiment was quantified by Tapestation analysis. 600 pg of cDNA was tagmented by Nextera XT with the custom primers, P5_TSO_Hybrid and Nextera 70X. The single-cell sequencing library from each batch was either pooled with another batch or sequenced separately on the Illumina HiSeq2500 at 1.4pM or MiSeq at 8pM, with custom priming (Read1CustSeqB Drop-seq primer).

Histological methods

Collection and processing of testes

For histological studies, testes were collected in 4% paraformaldehyde (PFA), incubated overnight at 4˚C and washed with 70% ethanol. For hematoxylin and eosin staining, testes were collected in modified Davidson fixative and after 24 hr incubation at room temperature, tissues were transferred to Bouin's solution for another 24 hr incubation at room temperature. Fixed testes were dehydrated through a series of graded ethanol baths and embedded in paraffin. Then 5 µm sections were cut on clean glass slides.

Hematoxylin and Eosin (HE) Staining

Hematoxylin and Eosin staining was performed on each mouse line (Wildtype, *Mlh3*$^{-/-}$, *Hormad1*$^{-/-}$, *Cul4a*$^{-/-}$, and *CNP-EGFP*) to assess overall morphology of testicular tissue. Slides were deparaffinized with xylene and rehydrated through a series of graded ethanol bath to PBS. Standard HE staining protocol was adapted from Belinda Dana (Department of Ophthalmology, Washington University in St. Louis) and followed with Hematoxylin 560% and 1% Alcoholic Eosin Y 515.

Immunofluorescence staining

Prior to immunofluorescence staining, antigen retrieval was performed by boiling slides in citric acid buffer for 20 min, and tissue sections were blocked in blocking solution (0.5% Triton X-100 +2% goat serum in 1X PBS) for an hour at room temperature. Primary antibodies were diluted to antibody-specific dilution (see Key Resources Table) and incubated overnight at 4˚C in a humid chamber. Then, slides were incubated in secondary antibodies (1:300 dilution) at room temperature for 4 hr in a humid chamber. After the secondary antibody incubation, sections were stained with Hoechst (1:500 dilution), washed with 1X PBS and mounted with ProLong Diamond Antifade Mountant for visualization under confocal microscope.

Computational methods

Preprocessing of Drop-seq data

Paired-end sequencing reads were processed, filtered and aligned as described in ***Macosko et al. (2015)***. The specific steps and tools for this process is further outlined in Drop-seq Computational Cookbook (http://mccarrolllab.com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf). STAR aligner was used to map the processed reads to mouse genome (***Dobin et al., 2013***). A STAR indexed genome was generated using mm10 mouse genome and GRCm38 gene annotation (release version 76) with default setting. Following the alignment, digital gene expression (DGE) matrices were generated for each experimental batch (***Macosko et al., 2015***).

Quality control for Drop-seq data

Combined raw DGEs were processed through a series of quality control and normalization steps. Cells with fewer than 200 UMI counts or fewer than 50 genes expressed were removed. Cells were also removed if their total UMI count or number of genes expressed was more than one standard deviation below the mean for that experiment. A t-SNE reduction of this dataset revealed an amorphous homogeneous group characterized by a low library size, high mitochondrial gene expression and often co-expressed genes from early and late meiosis suggesting poor quality and or doublet cells and so these were removed. Cells with a normalized mt-Rnr2 expression of greater than two were also removed. After these steps 20,322 cells and 28,893 genes remained.

Genes in the lower third of expression means were then removed and cells were normalized by square root transformation of total transcript counts per cell and genes were normalized to unit variance. All expression values were capped to maximum of 10. This results in a final matrix of 20,322 cells by 19,262 genes with a sparsity of 93.8% and a median UMI count of 1312 per cell.

K-means clustering and differential expression analysis

K-means clustering was performed on the t-SNE result of SDA run (the version that removed likely components that represent batch effects and technical artifacts) using 'kmeans' in R, testing different numbers for 'k' (***Figure 1—figure supplement 2***) with maximum iterations set to 10,000. We first settled with 'k = 42 which slightly over-clustered the data (i.e. created more clusters than necessary) and then merged clusters that are transcriptionally indistinguishable. Briefly, a classification hierarchy tree that places transcriptionally similar clusters together was built using BuildClusterTree() function in Seurat(v2.3.0). To test for which clusters to be merged, the out-of-bag error (OOBE) method from a random forest classifier was used (implemented in Seurat via AssessNodes() and MergeNode() functions). The classification error was computed for left or right cells on each node of the tree and top five nodes with high OOBE were merged to finally produce 32 clusters in 'merged' t-SNE plot in ***Figure 1—figure supplement 2***. Then, differentially expressed markers for all k-means clusters were identified using FindAllMarkers() function in Seurat with 'min.pct' parameter set to '0.25' where genes that are detected in a minimum fraction of 0.25 cells will be tested for differentially expressed genes. This differentially expressed genes list was used for assigning cell-types to each k-mean cluster and generating a list of potential novel cell-type specific markers by extracting top 10 differentially expressed genes for each cell-type and removing genes that were already annotated in the literature. A selected number of markers on this list was validated using immunofluorescence.

Somatic cell population heterogeneity analysis using seurat

Seurat (v2.3.0) was used to subset, re-cluster and visualize somatic cell population data from joint wild-type and mutant dataset. After subsetting somatic cells from the original k-means result joint data (clusters 1, 2, 3, 4, 5, 8 and nine from ***Figure 1—figure supplement 2A***), the percentage of mitochondrial genes was re-calculated and then a linear transformation was applied (using ScaleData () function in Seurat) while regressing out unwanted source of variations (percentage of mitochondrial genes, number of transcripts, number of genes and batch). PCA was performed on the scaled data to reduce the dimensionality of the data. A number of statistically significant principle components (PCs) for clustering purpose was determined by plotting and examining the variability explained by each PC in decreasing order (using PCElbowPlot() function in Seurat). For clustering somatic cells, we used PC = 18 as an input for K-nearest neighbor (KNN) graph-based algorithm implemented in Seurat (FindClusters()) along with resolution parameter set to '0.5.' We used t-SNE to visualize the data and clustering result. Differentially expressed genes (DEGs) were identified using Seurat's FindAllMarkers() function with 'min.pct' parameter set to '0.25' where genes that are detected in a minimum fraction of 0.25 cells are tested for DEGs. The DEGs list was used for performing gene ontology enrichment analysis to retrieve a functional profile for each somatic cluster. p-values were corrected using Benjamini-Hochberg.

Sparse Decomposition of Arrays (SDA)

SDA v1.1 (***Hore, 2015***; ***Hore et al., 2016***) was then run on the post-QC final matrix with 50 components for 10,000 iterations to confirm convergence (although in practice the results are almost identical after just 1000 iterations). The number of components was chosen such that there were typically between 1 and 5 single cell components across runs. Briefly, SDA decomposes a DGE into a number of components represented by two matrices. The columns vectors of the first matrix indicate how much a given component is active in each cell and the rows of the second matrix indicate which genes are active in a given component. SDA convergence was confirmed using the change in free energy, as well as the change in fraction of posterior inclusion probabilities (PIPs, probability that a gene loading is not equal to zero i.e. not in the spike) less than 0.5. The distribution of PIPs, cell scores, and gene loadings were also assessed. SDA was also run four further times with different seeds as well as with different number of components to verify stability of the results. Those components with a single high loading in one cell (1, 4, 14, 18, 46) were removed to visualize relationships between the components. To visualize and quantify the biological relationships among cells, t-SNE (without initial PCA step) was run on a version of the component scores matrix with likely technical artifacts and batch components removed, using a 'perplexity' parameter of 50, and 1000 iterations (Rtsne package; ***Krijthe, 2015***; ***der and Hinton, 2008***; ***van der Maaten, 2014***). Technical components were manually identified as meeting one or both of the following criteria: two batches of the same mouse line had opposite or very different cell scores (components 6, 12, 22, 28, 29, 41) and or if the highest loading genes were all or mostly ribosomal or pseudogenes (components 9, 25, 43). To assess uncertainty in the t-SNE embedding, t-SNE was also run multiple times with different seeds (***Figure 3—figure supplement 5***). We also performed dimensionality reduction using UMAP and confirmed that it gave a pseudotime embedding consistent with t-SNE (***McInnes et al., 2018***) (***Figure 3—figure supplement 5***).

Note that SDA components have arbitrary sign and must be interpreted through the combination of gene and cell signs. Gene loadings and cell scores with concordant signs results in a positive expression contribution from a component, whereas discordant signs results in negative contribution.

To generate a pseudo-timeline we used a similar approach to that implemented in SCUBA (***Marco et al., 2014***). We iteratively fit a principal curve through the t-SNE plot with increasing degrees of freedom from 4 to 9, using the curve from the previous run as the starting point using the princurve package in R (***Hastie and Stuetzle, 1989***). Each cell was then assigned to its closest position on this curve, to define pseudotime for that cell. Somatic cells and the *Hormad1*$^{-/-}$ X-activated cells (component 38 score >3) were excluded during pseudotime construction, but the *Hormad1*$^{-/-}$ X-activated cells were given pseudotimes *post-hoc*. Somatic cells were defined by thresholding the cell scores of somatic components (if the absolute cell score of a given cell passed

any of the following component thresholds 26, 11, 3, 32, 45, 24 > 2; 37 > 1.5; 40 > 1; or mt-Rnr2 expression >3).

The temporal order of components was determined by using a weighted mean of the pseudotime values, where the weights are the cell scores of the component. In addition, only those cells with an absolute cell score of greater than two contribute to the mean. To calculate simulated haploid non-sharing (***Figure 6***) cells with pseudotime > −10000 were randomly split into two groups. The predicted X expression was calculated as Original X * PseudoTime/10000 + e, where is is a random normal error with mean 0 and s.d. of 3.

Computational analysis was performed using R (***R Development Core Team, 2018***). Gene ontology enrichment analysis was performed on the top 250 genes from each component (from each side) using the enrichGO function from the clusterProfiler R package in which p-values are calculated based on the hypergeometric distribution and corrected for testing of multiple biological process GO terms using the Benjamini-Hochberg procedure (***Yu et al., 2012***). Plots were created using the ggplot2 package and extensions ggrepel, ggforce, ggseqlogo, ggnewscale, ggrastr, RColorBrewer, viridis, and cowplot (***Campitelli, 2019***; ***Garnier, 2018***; ***Neuwirth, 2014***; ***Pedersen, 2016***; ***Petukhov, 2018***; ***Wagih, 2017***; ***Wickham, 2016***; ***Wilke, 2018***). In addition the following R packages were used: data.table, Matrix (for sparse large matrix computations), biomaRt (for gene identifiers), shiny and shinycssloaders (for creating the interactive web application), ComplexHeatmap, bigmemory (for creating a low-memory shiny app), and MASS (for kernel density estimation) (***Bates and Maechler, 2018***; ***Chang et al., 2018***; ***Dowle and Srinivasan, 2019***; ***Durinck et al., 2005***; ***Gu et al., 2016***; ***Kane et al., 2013***; ***Sali, 2017***; ***Venables and Ripley, 2002***).

Components were clustered by t-SNE, using either the absolute gene loadings or cell scores matrix (t-SNE perplexity = 2). Component names were then assigned based on known maker genes from the literature and cross checked for consistency against the distribution of components in t-SNE space. Components representing batch effects were identified by plotting cell scores by experimental batch and checking for biological subgroups with opposing cell scores.

We also ensured the KO cells were not unduly affecting the estimated components by separately performing an SDA analysis with only WT cells (normalized separately but with the same parameters). The same number of iterations, number of components, and random seed, were used. To account for rotations of the results we performed a procrustean rotation on the WT loadings matrix with the mixed loadings matrix as the target. Procrustes rotation was performed using the R package vegan (***Lin and Boutros, 2019***; ***Oksanen et al., 2019***). We correlated the gene loadings of the Mixed WT and KO SDA analysis with the WT only analysis (after rotation) and found strong correspondence for those WT components which contained many cells (***Figure 3—figure supplements 2–4***).

Validation of SDA imputation

Imputed gene expression values (the posterior means of the SDA model) were computed as the matrix product of the cell scores and gene loadings matrix from SDA.

In order to formally quantify the accuracy of SDA imputation, we performed a cross validation study comparing the ability of SDA imputation to correctly predict single cell gene expression data in a withheld sample. First, we randomly split the post-QC RNA-sequencing reads from the full dataset into two batches: with 20% probability a read is assigned to the test dataset, and with 80% probability it is assigned to the training dataset. Next we create seven predictors of gene expression levels for each cell, using the training dataset: 'Unimputed' uses the training data directly (scaled by the total UMI counts for each cell), 'Mean cell' uses the sum of training reads for each gene across all cells to predict ranks (i.e. every cell has the same prediction), the matrix factorization approaches SDA, ICA, PCA and NNMF were run on the normalized training data (normalized as described above) and imputed values calculated as the matrix product of cell scores and gene loadings, MAGIC values were computed using the Rmagic package.

To compare the accuracy of the three predictors for gene expression imputation, we evaluate an objective function for each predictor and each cell, which we call the 'rank prediction accuracy curve' or RPAC. The RPAC for each predictor is created by rank ordering all genes in a single cell by the predicted level of expression of those genes, from high-to-low, after reversing normalisations (***Figure 5***). For each rank (abscissa), the ordinate is the cumulative fraction of test data reads for all

genes up to that rank (i.e. all genes with higher predicted expression than the current rank). The RPAC is similar in spirit to a receiver operating characteristic (ROC) curve. The area under the curve (AUC) for each RPAC is informative about prediction accuracy; a completely random predictor is expected to produce an AUC of 0.5, while a method with some predictive utility will have an AUC >0.5. This allows us to prefer predictions with a higher AUC, although we note that (unlike for a ROC) even given perfect imputation, the maximum possible expected AUC is <1, because the test data is sparse and so shows considerable noise relative to the unknown truth.

In order to identify differences between SDA and NNMF (the most similar alternative method), for each gene we calculated imputed expression for both methods (not using single cell components from SDA), and calculated Pearson correlation between the two methods. We then looked for enrichment (by FET, p=0.05 after correction for multiple testing by Bonferroni) of the 500 least correlated genes in both SDA and NNMF components, finding seven enriched SDA components (3P, 16N, 4N, 10P, 50N, 46N, 8P) and 0 NNMF components. We show example genes from 3P and 50N in ***Figure 5D and F***.

NNMF analysis was performed using the NNLM R package (***Lin and Boutros, 2019***) with 50 components, and a stop criterion of 10^{-5}. ICA analysis was performed with the fastICA R package (***Marchini et al., 2017***) with 50 components. PCA was performed using the R package flashpcaR with 50 components and divisor and standardization set to 'none' (***Abraham and Inouye, 2014***). MAGIC was performed with default parameters using the R package Rmagic 1.5.0. (***van Dijk et al., 2018***).

Transcription factor motif analysis

To discover de novo motifs enriched within each component we used the MotifFinder software - an iterative Gibbs sampler described in ***Altemose et al. (2017)***; ***Davies et al. (2016)***. We ran the MotifFinder R package on the repeat masked promoter sequences from *Mus Musculus* GRCm38 of the top 250 positive and negative genes (separately) for each component. Sequences with greater than 10% masked bases were removed. For each component 10 different regions around the TSS were used (150, 200, 250 bp upstream and downstream of the TSS (separately), and 200, 300, 400, and 800 bp centered on the TSS). Each run was repeated nine times with different random seed motifs (each of length 6 bp). In each case MotifFinder was run for 1000 iterations or until convergence (defined as when standard deviation of the motif proportion is below 0.05).

The resulting de novo motifs were annotated with known motifs from the HOCOMOCO database (V11) using Tomtom from the MEME suite (***Bailey et al., 2015***; ***Gupta et al., 2007***; ***Kulakovskiy et al., 2018***). We subset the matches by taking the match with the minimum q-value for each HOCOMOCO target for those with a E-value of less than 0.001. This resulted in 123 different matches from 101 de novo motifs. These were manually grouped into 16 categories based on motif similarity. Within each group a single 'most likely acting' motif was chosen based on external suggestive evidence, including whether a knockout of the transcription factor (TF) causes infertility, if the TF is specifically expressed in testis by RNA-Seq data in GTEx (GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2)), or if the TF is previously known to bind testis expressed genes by ChIP-Seq (***Lonsdale et al., 2013***).

To find good motifs with poor matches to currently known motifs we plotted the sum of the information content by the E value for each de novo motif. This identified a set of similar motifs with poor E values but large total information content – most of these matched best to ATF1.

In order to determine patterns of association of motifs with pseudotime, we performed correlation tests (using R's cor.test) between the gene loadings of each component (positive and negative loadings separately) and the probability of the motif being present at the promoter of these genes. To find probability of motif presence per gene promoter, we used MotifFinder but fixed the position weight matrix to either one of our de novo motifs (***Figure 7***) or a motif from the HOCOMOCO database (***Figure 7—figure supplement 2***) and ran MotifFinder for 20 iterations. We assessed convergence through the change in proportion of sequences containing the motif.

We used published ChIP-Seq data for the transcription factors *Stra8* (***Kojima et al., 2019***), *Mybl1* (***Bolcun-Filas et al., 2011***), *Crem* (***Kosir et al., 2012***), and *Rfx2* (***Kistler et al., 2015***) to validate our conclusion that some SDA components correspond to genes coregulated by these transcription factors. We performed a Fisher's exact test on the overlap of the genes suggested from the ChIP-Seq

studies for each of the three transcription factors with the top 500 genes for each SDA component (positive and negative loadings separately). For *Stra8* the genes are those described as 'STRA8-activated' from supplementary table 1 of Kojima et al. For *Mybl1* the genes are those determined as potential direct targets of MYBL1 (those that were bound in ChIP and were mis-regulated in *repro9* mice, from Table 1 of Bolcun-Filas et al.). For *Crem* the genes were those that are "found in the cross-section between DE genes from Kosir et al. and genes bound by CREM in testis from Martianov et al' (i.e. the top 50 genes from Kosir, et al. Supp Table 4). For *Rfx2* the genes are the genes from Kistler et al. that are bound by RFX2 by ChIP-seq and are statistically significantly downregulated at P30 (from their Table S1).

Component enrichment analysis

The goal of this analysis (shown in ***Figure 8B***) is to simply assess whether cell scores for each component are randomly distributed across strains. In our straightforward approach, we assess the most hypothesis-free null expectation: that cell component loadings are completely randomly distributed across cells, regardless of cell type or mouse strain. We use the full set of cells assessed by SDA. Each strain, i, contributed p_i proportion of cells to the total SDA dataset, represented by the dashed horizontal line plotted for each strain. Under our most naive assumption, the proportion of cell from strain i that load on component j, f_{ij}, should be equal to p_i. If $f_{ij} > p_i$ we say that the component j is enriched in strain i.

Apolipoprotein E (ApoE) Immunofluorescence Signal Quantification

To quantify the frequency of ApoE protein signal in wild-type and mutant animals, we counted the total number of intact testicular tubules present on slides and the number of tubules with ApoE protein signal using a confocal microscope at 20x. A Fisher's exact test was used to test the hypothesis that the frequency of ApoE-positive tubules was the same in wild-type, *Mlh3*$^{-/-}$ and *Hormad1*$^{-/-}$ strains.

Data and software availability

Raw data and processed files for Drop-seq experiments are available under GEO accession number GSE113293.

R markdown files that enable simulating main steps of the analysis are available upon reasonable request. Custom R code used is available at www.github.com/MyersGroup/testisAtlas (***Wells, 2019***) and archived at DOI: 10.5281/zenodo.3233958.

SDA is available from https://jmarchini.org/sda/.

Acknowledgements

We thank Abul Usmani for assistance with mouse husbandry and advice on *Pou5f1*:GFP reporter animals, Jeffrey Milbrandt and the WashU Genetics Department Single Cell Program for support, Liang Ma for providing *Cul4a*$^{-/-}$ mice, Joe Dougherty for providing *Cnp* mice, and Katinka Vigh-Conrad for assistance with figures. We also thank the Alvin J Siteman Cancer Center at Washington University School of Medicine and Barnes-Jewish Hospital in St. Louis, MO, for the use of the High-Speed Cell Sorter Core, which provided cell sorting service. The Siteman Cancer Center is supported in part by an NCI Cancer Center Support Grant #P30 CA91842. This work was supported by National Institutes of Health Grants R01HD078641 and R01MH101810 to DFC, and Wellcome Trust grants 098387/Z/12/Z and 212284/Z/18/Z to SM and 109109/Z/15/Z to DW. Research reported in this publication was supported by the Office of the Director, of the National Institutes of Health under Award Number P51OD011092 to the Oregon National Primate Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Additional information

Funding

Funder	Grant reference number	Author
Eunice Kennedy Shriver National Institute of Child Health and Human Development	R01HD078641	Donald F Conrad
National Institute of Mental Health	R01MH101810	Donald F Conrad
Wellcome	098387/Z/12/Z	Simon R Myers
Wellcome	109109/Z/15/Z	Daniel Wells
European Research Council	617306	Jonathan Marchini
Wellcome	212284/Z/18/Z	Simon R Myers

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Min Jung, Formal analysis, Investigation, Visualization, Methodology, Writing—original draft; Daniel Wells, Conceptualization, Data curation, Software, Formal analysis, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Jannette Rusch, Supervision, Investigation, Visualization; Suhaira Ahmad, Investigation; Jonathan Marchini, Conceptualization, Software, Supervision, Funding acquisition, Methodology, Writing—review and editing; Simon R Myers, Conceptualization, Formal analysis, Supervision, Funding acquisition, Methodology, Project administration, Writing—review and editing; Donald F Conrad, Conceptualization, Resources, Supervision, Funding acquisition, Investigation, Methodology, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Daniel Wells https://orcid.org/0000-0002-2007-8978
Jonathan Marchini https://orcid.org/0000-0003-0610-8322
Donald F Conrad https://orcid.org/0000-0003-3828-8970

Ethics

Animal experimentation: All animal experiments were performed in compliance with the regulations of the Animal Studies Committee at Washington University in St. Louis under approved protocol #20160089.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.43966.040
Author response https://doi.org/10.7554/eLife.43966.041

Additional files

Supplementary files

- Supplementary file 1. Summary of all wild-type and mutant single-cell RNA-sequencing experiments.
DOI: https://doi.org/10.7554/eLife.43966.029
- Supplementary file 2. Summary of all differentially expressed genes in total joint wild-type and mutant cell clusters.
DOI: https://doi.org/10.7554/eLife.43966.030
- Supplementary file 3. Component overview. A table of key genes from 26 example components.
DOI: https://doi.org/10.7554/eLife.43966.031

• Supplementary file 4. A ZIP file containing results of the full SDA analysis reported in the manuscript, which can be loaded and explored in the R computing environment.
DOI: https://doi.org/10.7554/eLife.43966.032

• Supplementary file 5. A ZIP file containing all the de novo inferred motifs in MEME format, in addition to tables summarizing the best tomtom HOCOMOCO matches for each of these.
DOI: https://doi.org/10.7554/eLife.43966.033

• Supplementary file 6. GO Categories related to amyloid-beta metabolism show significant enrichment in components 49, 26 and 16.
DOI: https://doi.org/10.7554/eLife.43966.034

• Supplementary file 7. Summary of SDA runtime and memory usage for example datasets.
DOI: https://doi.org/10.7554/eLife.43966.035

• Transparent reporting form
DOI: https://doi.org/10.7554/eLife.43966.036

Data availability

Raw data and processed files for Drop-seq and 10X Genomics experiments are available in GEO under accession number GSE113293.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Jung M, Wells, DJ, Rusch J, Ahmad S, Marchini J, Myers S, Conrad DF	2019	A single-cell atlas of testis gene expression from 5 mouse strains	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113293	NCBI Gene Expression Omnibus, GSE113293

References

Abby E, Tourpin S, Ribeiro J, Daniel K, Messiaen S, Moison D, Guerquin J, Gaillard JC, Armengaud J, Langa F, Toth A, Martini E, Livera G. 2016. Implementation of meiosis prophase I programme requires a conserved retinoid-independent stabilizer of meiotic transcripts. *Nature Communications* **7**:10324. DOI: https://doi.org/10.1038/ncomms10324, PMID: 26742488

Abraham G, Inouye M. 2014. Fast principal component analysis of large-scale genome-wide data. *PLOS ONE* **9**: e93766. DOI: https://doi.org/10.1371/journal.pone.0093766, PMID: 24718290

Adelman CA, Petrini JH. 2008. ZIP4H (TEX11) deficiency in the mouse impairs meiotic double strand break repair and the regulation of crossing over. *PLOS Genetics* **4**:e1000042. DOI: https://doi.org/10.1371/journal.pgen.1000042, PMID: 18369460

Altemose N, Noor N, Bitoun E, Tumian A, Imbeault M, Chapman JR, Aricescu AR, Myers SR. 2017. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *eLife* **6**:e28383. DOI: https://doi.org/10.7554/eLife.28383, PMID: 29072575

Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* **97**:10101–10106. DOI: https://doi.org/10.1073/pnas.97.18.10101, PMID: 10963673

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME suite. *Nucleic Acids Research* **43**:W39–W49. DOI: https://doi.org/10.1093/nar/gkv416, PMID: 25953851

Banito A, Li X, Laporte AN, Roe J-S, Sanchez-Vega F, Huang C-H, Dancsok AR, Hatzi K, Chen C-C, Tschaharganeh DF, Chandwani R, Tasdemir N, Jones KB, Capecchi MR, Vakoc CR, Schultz N, Ladanyi M, Nielsen TO, Lowe SW. 2018. The SS18-SSX oncoprotein hijacks KDM2B-PRC1.1 to Drive Synovial Sarcoma. *Cancer Cell* **33**:527–541. DOI: https://doi.org/10.1016/j.ccell.2018.01.018

Bates D, Maechler M. 2018. Matrix: Sparse and Dense Matrix Classes and Methods.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**:836–840. DOI: https://doi.org/10.1126/science.1183439, PMID: 20044539

Bernardo JM. 2003. *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*. Oxford University Press.

Boateng KA, Bellani MA, Gregoretti IV, Pratto F, Camerini-Otero RD. 2013. Homologous pairing preceding SPO11-mediated double-strand breaks in mice. *Developmental Cell* **24**:196–205. DOI: https://doi.org/10.1016/j.devcel.2012.12.002, PMID: 23318132

Boekhout M, Karasu ME, Wang J, Acquaviva L, Pratto F, Brick K, Eng DY, Xu J, Camerini-Otero RD, Patel DJ, Keeney S. 2019. REC114 partner ANKRD31 controls number, timing, and location of meiotic DNA breaks. *Molecular Cell* **74**:1053–1068. DOI: https://doi.org/10.1016/j.molcel.2019.03.023, PMID: 31003867

Bolcun-Filas E, Bannister LA, Barash A, Schimenti KJ, Hartford SA, Eppig JJ, Handel MA, Shen L, Schimenti JC. 2011. A-MYB (MYBL1) transcription factor is a master regulator of male meiosis. *Development* **138**:3319–3330. DOI: https://doi.org/10.1242/dev.067645, PMID: 21750041

Braun RE, Behringer RR, Peschon JJ, Brinster RL, Palmiter RD. 1989. Genetically haploid spermatids are phenotypically diploid. *Nature* **337**:373–376. DOI: https://doi.org/10.1038/337373a0

Brown MS, Grubb J, Zhang A, Rust MJ, Bishop DK. 2015. Small Rad51 and Dmc1 complexes often Co-occupy both ends of a meiotic DNA double strand break. *PLOS Genetics* **11**:e1005653. DOI: https://doi.org/10.1371/journal.pgen.1005653, PMID: 26719980

Brown MS, Bishop DK. 2015. DNA strand exchange and RecA homologs in meiosis. *Cold Spring Harbor Perspectives in Biology* **7**:a016659. DOI: https://doi.org/10.1101/cshperspect.a016659

Brunet JP, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *PNAS* **101**:4164–4169. DOI: https://doi.org/10.1073/pnas.0308531101, PMID: 15016911

Buaas FW, Kirsh AL, Sharma M, McLean DJ, Morris JL, Griswold MD, de Rooij DG, Braun RE. 2004. Plzf is required in adult male germ cells for stem cell self-renewal. *Nature Genetics* **36**:647–652. DOI: https://doi.org/10.1038/ng1366, PMID: 15156142

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**:411–420. DOI: https://doi.org/10.1038/nbt.4096, PMID: 29608179

Campitelli E. 2019. ggnewscale: Multiple Fill and Color Scales in 'ggplot2'.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. 2018. shiny: Web Application Framework for R.

Chen Y, Zheng Y, Gao Y, Lin Z, Yang S, Wang T, Wang Q, Xie N, Hua R, Liu M, Sha J, Griswold MD, Li J, Tang F, Tong M-H. 2018. Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Research* **28**:879–896. DOI: https://doi.org/10.1038/s41422-018-0074-y

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, Hlavina W, Kapustin Y, Meric P, Maglott D, Birtle Z, Marques AC, Graves T, Zhou S, Teague B, Potamousis K, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLOS Biology* **7**:e1000112. DOI: https://doi.org/10.1371/journal.pbio.1000112, PMID: 19468303

da Cruz I, Rodríguez-Casuriaga R, Santiñaque FF, Farías J, Curti G, Capoano CA, Folle GA, Benavente R, Sotelo-Silveira JR, Geisinger A. 2016. Transcriptome analysis of highly purified mouse spermatogenic cell populations: gene expression signatures switch from meiotic-to postmeiotic-related processes at pachytene stage. *BMC Genomics* **17**:294. DOI: https://doi.org/10.1186/s12864-016-2618-1, PMID: 27094866

Dai J, Voloshin O, Potapova S, Camerini-Otero RD. 2017. Meiotic knockdown and complementation reveals essential role of RAD51 in mouse spermatogenesis. *Cell Reports* **18**:1383–1394. DOI: https://doi.org/10.1016/j.celrep.2017.01.024

Daniel K, Lange J, Hached K, Fu J, Anastassiadis K, Roig I, Cooke HJ, Stewart AF, Wassmann K, Jasin M, Keeney S, Tóth A. 2011. Meiotic homologue alignment and its quality surveillance are controlled by mouse HORMAD1. *Nature Cell Biology* **13**:599–610. DOI: https://doi.org/10.1038/ncb2213, PMID: 21478856

Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, Hinch AG, Moralli D, Biggs D, Diaz R, Preece C, Li R, Bitoun E, Brick K, Green CM, Camerini-Otero RD, Myers SR, Donnelly P. 2016. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* **530**:171–176. DOI: https://doi.org/10.1038/nature16931

der MLvan, Hinton G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research : JMLR* **9**:2579–2605.

Diagouraga B, Clément JAJ, Duret L, Kadlec J, de Massy B, Baudat F. 2018. PRDM9 methyltransferase activity is essential for meiotic DNA Double-Strand break formation at its binding sites. *Molecular Cell* **69**:853–865. DOI: https://doi.org/10.1016/j.molcel.2018.01.033, PMID: 29478809

Ding X, Xu R, Yu J, Xu T, Zhuang Y, Han M. 2007. SUN1 is required for telomere attachment to nuclear envelope and gametogenesis in mice. *Developmental Cell* **12**:863–872. DOI: https://doi.org/10.1016/j.devcel.2007.03.018, PMID: 17543860

Djureinovic D, Fagerberg L, Hallström B, Danielsson A, Lindskog C, Uhlén M, Pontén F. 2014. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *MHR: Basic Science of Reproductive Medicine* **20**:476–488. DOI: https://doi.org/10.1093/molehr/gau018

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. DOI: https://doi.org/10.1093/bioinformatics/bts635, PMID: 23104886

Dohle GR, Elzanaty S, van Casteren NJ. 2012. Testicular biopsy: clinical practice and interpretation. *Asian Journal of Andrology* **14**:88–93. DOI: https://doi.org/10.1038/aja.2011.57, PMID: 22157985

Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. 2015. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**:575–579. DOI: https://doi.org/10.1038/nature16462, PMID: 26675734

Dowle M, Srinivasan A. 2019. data.table: Extension of 'data.frame.

Dubiel W, Dubiel D, Wolf DA, Naumann M. 2018. Cullin 3-Based ubiquitin ligases as master regulators of mammalian cell differentiation. *Trends in Biochemical Sciences* **43**:95–107. DOI: https://doi.org/10.1016/j.tibs.2017.11.010, PMID: 29249570

Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, Wang Y, Wong WH. 2018. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *PNAS* **115**:7723–7728. DOI: https://doi.org/10.1073/pnas.1805681115, PMID: 29987051

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**:3439–3440. DOI: https://doi.org/10.1093/bioinformatics/bti525

Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. 2019. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* **10**:390. DOI: https://doi.org/10.1038/s41467-018-07931-2, PMID: 30674886

Ernst C, Eling N, Martinez-Jimenez CP, Marioni JC, Odom DT. 2019. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nature Communications* **10**:1251. DOI: https://doi.org/10.1038/s41467-019-09182-1, PMID: 30890697

Frungieri MB, Calandra RS, Lustig L, Meineke V, Köhn FM, Vogt HJ, Mayerhofer A. 2002. Number, distribution pattern, and identification of macrophages in the testes of infertile men. *Fertility and Sterility* **78**:298–306. DOI: https://doi.org/10.1016/S0015-0282(02)03206-5, PMID: 12137866

Garcia T, Hofmann MC. 2012. Isolation of undifferentiated and early differentiating type A spermatogonia from Pou5f1-GFP reporter mice. *Methods in Molecular Biology* **825**:31–44. DOI: https://doi.org/10.1007/978-1-61779-436-0_3, PMID: 22144234

Garnier S. 2018. viridis: Default Color Maps from matplotlib.

Geisinger A, Rodríguez-Casuriaga R. 2010. Flow cytometry for gene expression studies in mammalian spermatogenesis. *Cytogenetic and Genome Research* **128**:46–56. DOI: https://doi.org/10.1159/000291489, PMID: 20389037

Getun IV, Torres B, Bois PR. 2011. Flow cytometry purification of mouse meiotic cells. *Journal of Visualized Experiments*. DOI: https://doi.org/10.3791/2602, PMID: 21525843

Goertz MJ, Wu Z, Gallardo TD, Hamra FK, Castrillon DH. 2011. Foxo1 is required in mouse spermatogonial stem cells for their maintenance and the initiation of spermatogenesis. *Journal of Clinical Investigation* **121**:3456–3466. DOI: https://doi.org/10.1172/JCI57984, PMID: 21865646

Goluža T, Boscanin A, Cvetko J, Kozina V, Kosović M, Bernat MM, Kasum M, Kaštelan Željko, Ježek D. 2014. Macrophages and leydig cells in testicular biopsies of azoospermic men. *BioMed Research International* **2014**:1–14. DOI: https://doi.org/10.1155/2014/828697

Gómez-H L, Felipe-Medina N, Sánchez-Martín M, Davies OR, Ramos I, García-Tuñón I, de Rooij DG, Dereli I, Tóth A, Barbero JL, Benavente R, Llano E, Pendas AM. 2016. C14ORF39/SIX6OS1 is a constituent of the synaptonemal complex and is essential for mouse fertility. *Nature Communications* **7**:13298. DOI: https://doi.org/10.1038/ncomms13298, PMID: 27796301

Green CD, Ma Q, Manske GL, Shami AN, Zheng X, Marini S, Moritz L, Sultan C, Gurczynski SJ, Moore BB, Tallquist MD, Li JZ, Hammoud SS. 2018. A comprehensive roadmap of murine spermatogenesis defined by Single-Cell RNA-Seq. *Developmental Cell* **46**:651–667. DOI: https://doi.org/10.1016/j.devcel.2018.07.025, PMID: 30146481

Greenbaum MP, Iwamori T, Buchold GM, Matzuk MM. 2011. Germ cell intercellular bridges. *Cold Spring Harbor Perspectives in Biology* **3**:a005850. DOI: https://doi.org/10.1101/cshperspect.a005850, PMID: 21669984

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**:2847–2849. DOI: https://doi.org/10.1093/bioinformatics/btw313, PMID: 27207943

Guiraldelli MF, Eyster C, Wilkerson JL, Dresser ME, Pezza RJ. 2013. Mouse HFM1/Mer3 is required for crossover formation and complete synapsis of homologous chromosomes during meiosis. *PLOS Genetics* **9**:e1003383. DOI: https://doi.org/10.1371/journal.pgen.1003383, PMID: 23555294

Guiraldelli MF, Felberg A, Almeida LP, Parikh A, de Castro RO, Pezza RJ. 2018. SHOC1 is a ERCC4-(HhH)2-like protein, integral to the formation of crossover recombination intermediates during mammalian meiosis. *PLOS Genetics* **14**:e1007381. DOI: https://doi.org/10.1371/journal.pgen.1007381, PMID: 29742103

Guo JH, Huang Q, Studholme DJ, Wu CQ, Zhao Z. 2005. Transcriptomic analyses support the similarity of gene expression between brain and testis in human as well as mouse. *Cytogenetic and Genome Research* **111**:107–109. DOI: https://doi.org/10.1159/000086378, PMID: 16103650

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biology* **8**:R24. DOI: https://doi.org/10.1186/gb-2007-8-2-r24, PMID: 17324271

Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, Gudjonsson SA, Frigge ML, Thorleifsson G, Sigurdsson A, Stacey SN, Sulem P, Masson G, Helgason A, Gudbjartsson DF, Thorsteinsdottir U, et al. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**:eaau1043. DOI: https://doi.org/10.1126/science.aau1043, PMID: 30679340

Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, Cairns BR. 2009. Distinctive chromatin in human sperm packages genes for embryo development. *Nature* **460**:473–478. DOI: https://doi.org/10.1038/nature08162, PMID: 19525931

Hastie T, Stuetzle W. 1989. Principal curves. *Journal of the American Statistical Association* **84**:502–516. DOI: https://doi.org/10.1080/01621459.1989.10478797

Hauri S, Comoglio F, Seimiya M, Gerstung M, Glatter T, Hansen K, Aebersold R, Paro R, Gstaiger M, Beisel C. 2016. A High-Density map for navigating the human polycomb complexome. *Cell Reports* **17**:583–595. DOI: https://doi.org/10.1016/j.celrep.2016.08.096, PMID: 27705803

He Z, Jiang J, Hofmann MC, Dym M. 2007. Gfra1 silencing in mouse spermatogonial stem cells results in their differentiation via the inactivation of RET tyrosine kinase. *Biology of Reproduction* **77**:723–733. DOI: https://doi.org/10.1095/biolreprod.107.062513, PMID: 17625109

He F, Muto Y, Inoue M, Kigawa T, Shirouzu M, Terada T, Yokoyama S. 2010. Complex structure of the zf-CW domain and the H3K4me3 peptide. *RCSB Protein Data Bank*. . DOI: https://doi.org/10.2210/pdb2RR4/pdb

Hermann BP, Cheng K, Singh A, Roa-De La Cruz L, Mutoji KN, Chen IC, Gildersleeve H, Lehle JD, Mayo M, Westernströer B, Law NC, Oatley MJ, Velte EK, Niedenberger BA, Fritze D, Silber S, Geyer CB, Oatley JM, McCarrey JR. 2018. The mammalian spermatogenesis Single-Cell transcriptome, from spermatogonial stem cells to spermatids. *Cell Reports* **25**:1650–1667. DOI: https://doi.org/10.1016/j.celrep.2018.10.026, PMID: 30404016

Hess RA, de Franca LR. 2009. Spermatogenesis and cycle of the seminiferous EpitheliumAdvances in experimental medicine and biology. *Advances in Experimental Medicine and Biology* **639**:1–15.

Holstein AF. 1978. Spermatophagy in the seminiferous tubules and excurrent ducts of the testis in rhesus monkey and in man. *Andrologia* **10**:331–352. DOI: https://doi.org/10.1111/j.1439-0272.1978.tb03044.x, PMID: 102218

Hore V. 2015. Latent Variable Models for Analysing Multidimensional Gene Expression Data (DPhil). The University of Oxford. https://ora.ox.ac.uk/objects/uuid:ec62bc11-5c3f-467d-9ff3-f3c4eb29d140

Hore V, Viñuela A, Buil A, Knight J, McCarthy MI, Small K, Marchini J. 2016. Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics* **48**:1094–1100. DOI: https://doi.org/10.1038/ng.3624, PMID: 27479908

Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. 2016. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology* **17**:29. DOI: https://doi.org/10.1186/s13059-016-0888-1

Imai Y, Baudat F, Taillepierre M, Stanzione M, Toth A, de Massy B. 2017. The PRDM9 KRAB domain is required for meiosis and involved in protein interactions. *Chromosoma* **126**:681–695. DOI: https://doi.org/10.1007/s00412-017-0631-z

Ishiguro K, Kim J, Shibuya H, Hernández-Hernández A, Suzuki A, Fukagawa T, Shioi G, Kiyonari H, Li XC, Schimenti J, Höög C, Watanabe Y. 2014. Meiosis-specific cohesin mediates homolog recognition in mouse spermatocytes. *Genes & Development* **28**:594–607. DOI: https://doi.org/10.1101/gad.237313.113, PMID: 24589552

Ito C, Toshimori K. 2016. Acrosome markers of human sperm. *Anatomical Science International* **91**:128–142. DOI: https://doi.org/10.1007/s12565-015-0323-9

Kane MJ, Emerson J, Weston S. 2013. Scalable strategies for computing with massive data. *Journal of Statistical Software* **55**. DOI: https://doi.org/10.18637/jss.v055.i14

Kang HS, Chen LY, Lichti-Kaiser K, Liao G, Gerrish K, Bortner CD, Yao HH, Eddy EM, Jetten AM. 2016. Transcription factor GLIS3: a new and critical regulator of postnatal stages of mouse spermatogenesis. *Stem Cells* **34**:2772–2783. DOI: https://doi.org/10.1002/stem.2449, PMID: 27350140

Keeney S, Giroux CN, Kleckner N. 1997. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88**:375–384. DOI: https://doi.org/10.1016/S0092-8674(00)81876-0, PMID: 9039264

Kim PM, Tidor B. 2003. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research* **13**:1706–1718. DOI: https://doi.org/10.1101/gr.903503, PMID: 12840046

Kistler WS, Baas D, Lemeille S, Paschaki M, Seguin-Estevez Q, Barras E, Ma W, Duteyrat JL, Morlé L, Durand B, Reith W. 2015. RFX2 is a major transcriptional regulator of spermiogenesis. *PLOS Genetics* **11**:e1005368. DOI: https://doi.org/10.1371/journal.pgen.1005368, PMID: 26162102

Kojima ML, de Rooij DG, Page DC. 2019. Amplification of a broad transcriptional program by a common factor triggers the meiotic cell cycle in mice. *eLife* **8**:e43738. DOI: https://doi.org/10.7554/eLife.43738, PMID: 30810530

Kong A, Thorleifsson G, Frigge ML, Masson G, Gudbjartsson DF, Villemoes R, Magnusdottir E, Olafsdottir SB, Thorsteinsdottir U, Stefansson K. 2014. Common and low-frequency variants associated with genome-wide recombination rate. *Nature Genetics* **46**:11–16. DOI: https://doi.org/10.1038/ng.2833, PMID: 24270358

Kosir R, Juvan P, Perse M, Budefeld T, Majdic G, Fink M, Sassone-Corsi P, Rozman D. 2012. Novel insights into the downstream pathways and targets controlled by transcription factors CREM in the testis. *PLOS ONE* **7**:e31798. DOI: https://doi.org/10.1371/journal.pone.0031798, PMID: 22384077

Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, Sabeti PC. 2018. Identifying gene expression programs of cell-type identity and cellular activity with Single-Cell RNA-Seq. *bioRxiv*. DOI: https://doi.org/10.1101/310599

Kovalenko OV, Wiese C, Schild D. 2006. RAD51AP2, a novel vertebrate- and meiotic-specific protein, shares a conserved RAD51-interacting C-terminal domain with RAD51AP1/PIR51. *Nucleic Acids Research* **34**:5081–5092. DOI: https://doi.org/10.1093/nar/gkl665, PMID: 16990250

Krijthe JH. 2015. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation.

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, Kolpakov FA, Makeev VJ. 2018. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research* **46**:D252–D259. DOI: https://doi.org/10.1093/nar/gkx1106, PMID: 29140464

Kuroda N, Nakayama H, Miyazaki E, Hayashi Y, Toi M, Hiroi M, Enzan H. 2004. Distribution and role of CD34-positive stromal cells and myofibroblasts in human normal testicular stroma. *Histology and Histopathology* **19**:743–751. DOI: https://doi.org/10.14670/HH-19.743, PMID: 15168336

Lee KY, Im JS, Shibata E, Park J, Handa N, Kowalczykowski SC, Dutta A. 2015. MCM8-9 complex promotes resection of double-strand break ends by MRE11-RAD50-NBS1 complex. *Nature Communications* **6**:7744. DOI: https://doi.org/10.1038/ncomms8744, PMID: 26215093

Lima AC, Jung M, Rusch J, Usmani A, Lopes AM, Conrad DF. 2017. A standardized approach for multispecies purification of mammalian male germ cells by mechanical tissue dissociation and flow cytometry. *Journal of Visualized Experiments*. DOI: https://doi.org/10.3791/55913, PMID: 28745623

Lin X, Boutros PC. 2019. NNLM: Fast and Versatile Non-Negative Matrix Factorization.

Lipkin SM, Moens PB, Wang V, Lenzi M, Shanmugarajah D, Gilgeous A, Thomas J, Cheng J, Touchman JW, Green ED, Schwartzberg P, Collins FS, Cohen PE. 2002. Meiotic arrest and aneuploidy in MLH3-deficient mice. *Nature Genetics* **31**:385–390. DOI: https://doi.org/10.1038/ng931, PMID: 12091911

Löffler-Wirth H, Kalcher M, Binder H. 2015. oposSOM: r-package for high-dimensional portraying of genome-wide expression landscapes on bioconductor. *Bioinformatics* **31**:3225–3227. DOI: https://doi.org/10.1093/bioinformatics/btv342, PMID: 26063839

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, et al. 2013. The Genotype-Tissue expression (GTEx) project. *Nature Genetics* **45**:580–585. DOI: https://doi.org/10.1038/ng.2653, PMID: 23715323

Lukassen S, Bosch E, Ekici AB, Winterpacht A. 2018. Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Scientific Reports* **8**:6521. DOI: https://doi.org/10.1038/s41598-018-24725-0, PMID: 29695820

Lukaszewicz A, Lange J, Keeney S, Jasin M. 2018. Control of meiotic double-strand-break formation by ATM: local and global views. *Cell Cycle* **17**:1155–1172. DOI: https://doi.org/10.1080/15384101.2018.1464847, PMID: 29963942

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**:1202–1214. DOI: https://doi.org/10.1016/j.cell.2015.05.002

Marchini JL, Heaton C, Ripley BD. 2017. fastICA: FastICA Algorithms to Perform ICA and Projection Pursuit.

Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC. 2014. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *PNAS* **111**:E5643–E5650. DOI: https://doi.org/10.1073/pnas.1408993111, PMID: 25512504

Marini M, Rosa I, Guasti D, Gacci M, Sgambati E, Ibba-Manneschi L, Manetti M. 2018. Reappraising the microscopic anatomy of human testis: identification of telocyte networks in the peritubular and intertubular stromal space. *Scientific Reports* **8**:14780. DOI: https://doi.org/10.1038/s41598-018-33126-2, PMID: 30283023

Martin-DeLeon PA, Zhang H, Morales CR, Zhao Y, Rulon M, Barnoski BL, Chen H, Galileo DS. 2005. Spam1-associated transmission ratio distortion in mice: elucidating the mechanism. *Reproductive Biology and Endocrinology : RB&E* **3**:32. DOI: https://doi.org/10.1186/1477-7827-3-32

Martinez JS, von Nicolai C, Kim T, Ehlén Å, Mazin AV, Kowalczykowski SC, Carreira A. 2016. BRCA2 regulates DMC1-mediated recombination through the BRC repeats. *PNAS* **113**:3515–3520. DOI: https://doi.org/10.1073/pnas.1601691113, PMID: 26976601

McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: uniform manifold approximation and projection. *Journal of Open Source Software* **3**:861. DOI: https://doi.org/10.21105/joss.00861

Moretti C, Vaiman D, Tores F, Cocquet J. 2016. Expression and epigenomic landscape of the sex chromosomes in mouse post-meiotic male germ cells. *Epigenetics & Chromatin* **9**:47. DOI: https://doi.org/10.1186/s13072-016-0099-8, PMID: 27795737

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**:876–879. DOI: https://doi.org/10.1126/science.1182363, PMID: 20044541

Nantel F, Sassone-Corsi P. 1996. CREM: a transcriptional master switch during the spermatogenesis differentiation program. *Frontiers in Bioscience : A Journal and Virtual Library* **1**:266–269.

Neuwirth E. 2014. RColorBrewer: ColorBrewer Palettes.

Oakberg EF. 1956. Duration of spermatogenesis in the mouse and timing of stages of the cycle of the seminiferous epithelium. *American Journal of Anatomy* **99**:507–516. DOI: https://doi.org/10.1002/aja.1000990307, PMID: 13402729

Oakberg EF. 1957. Duration of spermatogenesis in the mouse. *Nature* **180**:1137–1138. DOI: https://doi.org/10.1038/1801137a0, PMID: 13483640

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2019. vegan: Community Ecology Package.

Ortega MM. 2016. Surveillance mechanisms in mammalian meiosis. Universitat Autònoma De Barcelona.

Otto SP, Scott MF, Immler S. 2015. Evolution of haploid selection in predominantly diploid organisms. *PNAS* **112**:15952–15957. DOI: https://doi.org/10.1073/pnas.1512004112, PMID: 26669442

Pacheco S, Maldonado-Linares A, Marcet-Ortega M, Rojas C, Martínez-Marchal A, Fuentes-Lazaro J, Lange J, Jasin M, Keeney S, Fernández-Capetillo O, Garcia-Caldés M, Roig I. 2018. ATR is required to complete meiotic recombination in mice. *Nature Communications* **9**:2622. DOI: https://doi.org/10.1038/s41467-018-04851-z, PMID: 29977027

Papanikos F, Clement JAJ, Testa E, Ravindranathan R, Grey C, Dereli I, Bondarieva A, Valerio-Cabrera S, Stanzione M, Schleiffer A, Jansa P, Lustyk D, Jifeng F, Forejt J, Barchi M, de Massy B, Toth A. 2018. ANKRD31

regulates spatiotemporal patterning of meiotic recombination initiation and ensures recombination between heterologous sex chromosomes in mice. *bioRxiv*. DOI: https://doi.org/10.1101/423293

Parvanov ED, Petkov PM, Paigen K. 2010. Prdm9 controls activation of mammalian recombination hotspots. *Science* **327**:835. DOI: https://doi.org/10.1126/science.1181495

Parvanov ED, Tian H, Billings T, Saxl RL, Spruce C, Aithal R, Krejci L, Paigen K, Petkov PM. 2017. PRDM9 interactions with other proteins provide a link between recombination hotspots and the chromosomal axis in meiosis. *Molecular Biology of the Cell* **28**:488–499. DOI: https://doi.org/10.1091/mbc.e16-09-0686, PMID: 27932493

Pedersen TL. 2016. ggforce: Accelerating "ggplot2.

Petukhov V. 2018. ggrastr: Raster layers for ggplot2.

Powers NR, Parvanov ED, Baker CL, Walker M, Petkov PM, Paigen K. 2016. The meiotic recombination activator PRDM9 trimethylates both H3K36 and H3K4 at recombination hotspots in vivo. *PLOS Genetics* **12**:e1006146. DOI: https://doi.org/10.1371/journal.pgen.1006146, PMID: 27362481

R Development Core Team. 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rakshambikai R, Srinivasan N, Nishant KT. 2013. Structural Insights into *Saccharomyces cerevisiae* Msh4–Msh5 Complex Function Using Homology Modeling. *PLOS ONE* **8**:e78753. DOI: https://doi.org/10.1371/journal.pone.0078753

Rankin S. 2015. Complex elaboration: making sense of meiotic cohesin dynamics. *FEBS Journal* **282**:2426–2443. DOI: https://doi.org/10.1111/febs.13301

Reinholdt LG, Schimenti JC. 2005. Mei1 is epistatic to Dmc1 during mouse meiosis. *Chromosoma* **114**:127–134. DOI: https://doi.org/10.1007/s00412-005-0346-4

Ribeiro J, Dupaigne P, Duquenne C, Veaute X, Petrillo C, Saintome C, Faklaris O, Busso D, Guerois R, Keeney S, Jain D, Martini E, Livera G. 2018. MEIOB and SPATA22 resemble RPA subunits and interact with the RPA complex to promote meiotic recombination. *bioRxiv*. DOI: https://doi.org/10.1101/358242

Robert T, Nore A, Brun C, Maffre C, Crimi B, Guichard V, Bourbon H-M, de Massy B. 2016. The TopoVIB-Like protein family is required for meiotic DNA double-strand break formation. *Science* **351**:943–949. DOI: https://doi.org/10.1126/science.aad5309

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139–140. DOI: https://doi.org/10.1093/bioinformatics/btp616, PMID: 19910308

Rona GB, Eleutherio ECA, Pinheiro AS. 2016. PWWP domains and their modes of sensing DNA and histone methylated lysines. *Biophysical Reviews* **8**:63–74. DOI: https://doi.org/10.1007/s12551-015-0190-6, PMID: 28510146

Rosa A, Ballarino M, Sorrentino A, Sthandier O, De Angelis FG, Marchioni M, Masella B, Guarini A, Fatica A, Peschle C, Bozzoni I. 2007. The interplay between the master transcription factor PU.1 and miR-424 regulates human monocyte/macrophage differentiation. *PNAS* **104**:19849–19854. DOI: https://doi.org/10.1073/pnas.0706963104, PMID: 18056638

Royo H, Polikiewicz G, Mahadevaiah SK, Prosser H, Mitchell M, Bradley A, de Rooij DG, Burgoyne PS, Turner JM. 2010. Evidence that meiotic sex chromosome inactivation is essential for male fertility. *Current Biology* **20**:2117–2123. DOI: https://doi.org/10.1016/j.cub.2010.11.010, PMID: 21093264

Sali A. 2017. shinycssloaders: Add CSS Loading Animations to "shiny" Outputs.

Sassone-Corsi P. 2000. CREM: a master-switch regulating the balance between differentiation and apoptosis in male germ cells. *Molecular Reproduction and Development* **56**:228–229. DOI: https://doi.org/10.1002/(SICI)1098-2795(200006)56:2+<228::AID-MRD2>3.0.CO;2-B, PMID: 10824972

Sassone-Corsi P. 2002. Unique chromatin remodeling and transcriptional regulation in spermatogenesis. *Science* **296**:2176–2178. DOI: https://doi.org/10.1126/science.1070963, PMID: 12077401

Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S, Goeva A, Nemesh J, Kamitaki N, Brumbaugh S, Kulp D, McCarroll SA. 2018. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**:1015–1030. DOI: https://doi.org/10.1016/j.cell.2018.07.028, PMID: 30096299

Schultz N, Hamra FK, Garbers DL. 2003. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. *PNAS* **100**:12201–12206. DOI: https://doi.org/10.1073/pnas.1635054100, PMID: 14526100

Shao C, Höfer T. 2017. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* **33**:235–242. DOI: https://doi.org/10.1093/bioinformatics/btw607, PMID: 27663498

Shin YH, Choi Y, Erdin SU, Yatsenko SA, Kloc M, Yang F, Wang PJ, Meistrich ML, Rajkovic A. 2010. Hormad1 mutation disrupts synaptonemal complex formation, recombination, and chromosome segregation in mammalian meiosis. *PLOS Genetics* **6**:e1001190. DOI: https://doi.org/10.1371/journal.pgen.1001190, PMID: 21079677

Sin HS, Kartashov AV, Hasegawa K, Barski A, Namekawa SH. 2015. Poised chromatin and bivalent domains facilitate the mitosis-to-meiosis transition in the male germline. *BMC Biology* **13**:53. DOI: https://doi.org/10.1186/s12915-015-0159-8, PMID: 26198001

Soh YQS, Mikedis MM, Kojima M, Godfrey AK, de Rooij DG, Page DC. 2017. Meioc maintains an extended meiotic prophase I in mice. *PLOS Genetics* **13**:e1006704. DOI: https://doi.org/10.1371/journal.pgen.1006704, PMID: 28380054

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, Dym M, de Massy B, Mikkelsen TS, Kaessmann H. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Reports* **3**:2179–2190. DOI: https://doi.org/10.1016/j.celrep.2013.05.031, PMID: 23791531

Stanzione M, Baumann M, Papanikos F, Dereli I, Lange J, Ramlal A, Tränkner D, Shibuya H, de Massy B, Watanabe Y, Jasin M, Keeney S, Tóth A. 2016. Meiotic DNA break formation requires the unsynapsed chromosome axis-binding protein IHO1 (CCDC36) in mice. *Nature Cell Biology* **18**:1208–1220. DOI: https://doi.org/10.1038/ncb3417, PMID: 27723721

Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, Xu Y, Fertig EJ. 2018. Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics* **34**:790–805. DOI: https://doi.org/10.1016/j.tig.2018.07.003, PMID: 30143323

Stojkov NJ, Janjic MM, Kostic TS, Andric SA. 2013. Orally applied doxazosin disturbed testosterone homeostasis and changed the transcriptional profile of steroidogenic machinery, cAMP/cGMP signalling and adrenergic receptors in leydig cells of adult rats. *Andrology* **1**:332–347. DOI: https://doi.org/10.1111/j.2047-2927.2012.00035.x, PMID: 23413145

Sun X, Brieño-Enríquez MA, Cornelius A, Modzelewski AJ, Maley TT, Campbell-Peterson KM, Holloway JK, Cohen PE. 2016. FancJ (Brip1) loss-of-function allele results in spermatogonial cell depletion during embryogenesis and altered processing of crossover sites during meiotic prophase I in mice. *Chromosoma* **125**:237–252. DOI: https://doi.org/10.1007/s00412-015-0549-2, PMID: 26490168

Suzuki H, Sada A, Yoshida S, Saga Y. 2009. The heterogeneity of spermatogonia is revealed by their topology and expression of marker proteins including the germ cell-specific proteins Nanos2 and Nanos3. *Developmental Biology* **336**:222–231. DOI: https://doi.org/10.1016/j.ydbio.2009.10.002, PMID: 19818747

Syrjänen JL, Pellegrini L, Davies OR. 2014. A molecular model for the role of SYCP3 in meiotic chromosome organisation. *eLife* **3**:e02963. DOI: https://doi.org/10.7554/eLife.02963

Tu Z, Bayazit MB, Liu H, Zhang J, Busayavalasa K, Risal S, Shao J, Satyanarayana A, Coppola V, Tessarollo L, Singh M, Zheng C, Han C, Chen Z, Kaldis P, Gustafsson JÅ, Liu K. 2017. Speedy A-Cdk2 binding mediates initial telomere-nuclear envelope attachment during meiotic prophase I independent of Cdk2 activation. *PNAS* **114**:592–597. DOI: https://doi.org/10.1073/pnas.1618465114, PMID: 28031483

Turner JM. 2007. Meiotic sex chromosome inactivation. *Development* **134**:1823–1831. DOI: https://doi.org/10.1242/dev.000018, PMID: 17329371

Turner JM. 2015. Meiotic silencing in mammals. *Annual Review of Genetics* **49**:395–412. DOI: https://doi.org/10.1146/annurev-genet-112414-055145, PMID: 26631513

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* **347**:1260419. DOI: https://doi.org/10.1126/science.1260419, PMID: 25613900

van der Maaten L. 2014. Accelerating t-SNE using Tree-Based algorithms. *Journal of Machine Learning Research : JMLR* **15**:3221–3245.

van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, Bierie B, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D. 2018. Recovering gene interactions from Single-Cell data using data diffusion. *Cell* **174**:716–729. DOI: https://doi.org/10.1016/j.cell.2018.05.061, PMID: 29961576

Venables WN, Ripley BD. 2002. Modern Applied Statistics with S.

Véron N, Bauer H, Weisse AY, Lüder G, Werber M, Herrmann BG. 2009. Retention of gene products in syncytial spermatids promotes non-Mendelian inheritance as revealed by the t complex responder. *Genes & Development* **23**:2705–2710. DOI: https://doi.org/10.1101/gad.553009, PMID: 19952105

Vrielynck N, Chambon A, Vezon D, Pereira L, Chelysheva L, De Muyt A, Mézard C, Mayer C, Grelon M. 2016. A DNA topoisomerase VI-like complex initiates meiotic recombination. *Science* **351**:939–943. DOI: https://doi.org/10.1126/science.aad5196, PMID: 26917763

Wagih O. 2017. Ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**:3645–3647. DOI: https://doi.org/10.1093/bioinformatics/btx469, PMID: 29036507

Wang W, Wei S, Li L, Su X, Du C, Li F, Geng B, Liu P, Xu G. 2015. Proteomic analysis of murine testes lipid droplets. *Scientific Reports* **5**:12070. DOI: https://doi.org/10.1038/srep12070, PMID: 26159641

Wang J, Tang C, Wang Q, Su J, Ni T, Yang W, Wang Y, Chen W, Liu X, Wang S, Zhang J, Song H, Zhu J, Wang Y. 2017. NRF1 coordinates with DNA methylation to regulate spermatogenesis. *The FASEB Journal* **31**:4959–4970. DOI: https://doi.org/10.1096/fj.201700093R, PMID: 28754714

Wang Y, Chen Y, Chen J, Wang L, Nie L, Long J, Chang H, Wu J, Huang C, Lei M. 2019. The meiotic TERB1-TERB2-MAJIN complex tethers telomeres to the nuclear envelope. *Nature Communications* **10**:564. DOI: https://doi.org/10.1038/s41467-019-08437-1, PMID: 30718482

Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. 2019. Single-Cell Multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**:1873–1887. DOI: https://doi.org/10.1016/j.cell.2019.05.006, PMID: 31178122

Wells D. 2019. testisAtlas. *GitHub*. https://github.com/MyersGroup/testisAtlas

Wickham H. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer.

Widger A, Mahadevaiah SK, Lange J, Ellnati E, Zohren J, Hirota T, Pacheco S, Maldonado-Linares A, Stanzione M, Ojarikre O, Maciulyte V, de Rooij DG, Tóth A, Roig I, Keeney S, Turner JMA. 2018. ATR is a multifunctional regulator of male mouse meiosis. *Nature Communications* **9**:2621. DOI: https://doi.org/10.1038/s41467-018-04850-0, PMID: 29976923

Wilke CO. 2018. cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2.
Wojtasz L, Daniel K, Roig I, Bolcun-Filas E, Xu H, Boonsanay V, Eckmann CR, Cooke HJ, Jasin M, Keeney S, McKay MJ, Toth A. 2009. Mouse HORMAD1 and HORMAD2, two conserved meiotic chromosomal proteins, are depleted from synapsed chromosome axes with the help of TRIP13 AAA-ATPase. *PLOS Genetics* **5**: e1000702. DOI: https://doi.org/10.1371/journal.pgen.1000702, PMID: 19851446
Xu Y, Greenberg RA, Schonbrunn E, Wang PJ. 2017. Meiosis-specific proteins MEIOB and SPATA22 cooperatively associate with the single-stranded DNA-binding replication protein A complex and DNA double-strand breaks. *Biology of Reproduction* **96**:1096–1104. DOI: https://doi.org/10.1093/biolre/iox040, PMID: 28453612
Yan W, McCarrey JR. 2009. Sex chromosome inactivation in the male. *Epigenetics* **4**:452–456. DOI: https://doi.org/10.4161/epi.4.7.9923, PMID: 19838052
Yang F, Eckardt S, Leu NA, McLaughlin KJ, Wang PJ. 2008. Mouse TEX15 is essential for DNA double-strand break repair and chromosomal synapsis during male meiosis. *The Journal of Cell Biology* **180**:673–679. DOI: https://doi.org/10.1083/jcb.200709057, PMID: 18283110
Yin Y, Lin C, Kim ST, Roig I, Chen H, Liu L, Veith GM, Jin RU, Keeney S, Jasin M, Moley K, Zhou P, Ma L. 2011. The E3 ubiquitin ligase cullin 4A regulates meiotic progression in mouse spermatogenesis. *Developmental Biology* **356**:51–62. DOI: https://doi.org/10.1016/j.ydbio.2011.05.661, PMID: 21624359
Yu G, Wang LG, Han Y, He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* **16**:284–287. DOI: https://doi.org/10.1089/omi.2011.0118, PMID: 22455463
Zhang J, Fujiwara Y, Yamamoto S, Shibuya H. 2019. A meiosis-specific BRCA2 binding protein recruits recombinases to DNA double-strand breaks to ensure homologous recombination. *Nature Communications* **10**:722. DOI: https://doi.org/10.1038/s41467-019-08676-2, PMID: 30760716
Zheng K, Wu X, Kaestner KH, Wang PJ. 2009. The pluripotency factor LIN28 marks undifferentiated spermatogonia in mouse. *BMC Developmental Biology* **9**:38. DOI: https://doi.org/10.1186/1471-213X-9-38, PMID: 19563657
Zhou Q, Wang M, Yuan Y, Wang X, Fu R, Wan H, Xie M, Liu M, Guo X, Zheng Y, Feng G, Shi Q, Zhao XY, Sha J, Zhou Q. 2016. Complete meiosis from embryonic stem Cell-Derived germ cells in vitro. *Cell Stem Cell*In **18**: 330–340. DOI: https://doi.org/10.1016/j.stem.2016.01.017, PMID: 26923202
Zhu X, Ching T, Pan X, Weissman SM, Garmire L. 2017. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ* **5**:e2888. DOI: https://doi.org/10.7717/peerj.2888, PMID: 28133571
Zickler D, Kleckner N. 2015. Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harbor Perspectives in Biology* **7**:a016626. DOI: https://doi.org/10.1101/cshperspect.a016626, PMID: 25986558